Applied Numerical Linear Algebra. Lecture 8

THEOREM *Let $A = U\Sigma V^T$ be the SVD of the m-by-n matrix A, where $m \geq n$. (There are analogous results for $m < n$.)*

- 1. Suppose that A is symmetric, with eigenvalues λ_i and orthonormal eigenvectors u_i. In other words $A = U\Lambda U^T$ is an eigendecomposition of A, with $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$, and $U = [u_1, \ldots, u_n]$, and $UU^T = I$. Then an SVD of A is $A = U\Sigma V^T$, where $\sigma_i = |\lambda_i|$ and $\upsilon_i = sign(\lambda_i)u_i$, where $sign(0) = 1$.
- 2. The eigenvalues of the symmetric matrix A^TA are σ_i^2. The right singular vectors υ_i are corresponding orthonormal eigenvectors.
- 3. The eigenvalues of the symmetric matrix AA^T are σ_i^2 and $m - n$ zeroes. The left singular vectors u_i are corresponding orthonormal eigenvectors for the eigenvalues σ_i^2. One can take any $m - n$ other orthogonal vectors as eigenvectors for the eigenvalue 0.
- 4. Let $H = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$, where A is square and $A = U\Sigma V^T$ is the SVD of A. Let $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$, $U = [u_1, \ldots, u_n]$, and $V = [\upsilon_1, \ldots, \upsilon_n]$. Then the 2n eigenvalues of H are $\pm\sigma_i$, with corresponding unit eigenvectors $\frac{1}{\sqrt{2}} \begin{bmatrix} \upsilon_i \\ \pm u_i \end{bmatrix}$.

- 5. If A has full rank, the solution of $\min_x \|Ax - b\|_2$ is $x = V\Sigma^{-1}U^T b$.
- 6. $\|A\|_2 = \sigma_1$. If A is square and nonsingular, then $\|A^{-1}\|_2^{-1} = \sigma_n$ and $\|A\|_2 \cdot \|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}$.
- 7. Write $V = [v_1, v_2, \ldots, v_n]$ and $U = [u_1, u_2, \ldots, u_n]$, so $A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T$ (a sum of rank-1 matrices). Then a matrix of rank $k < n$ closest to A (measured with $||\cdot||_2$) is $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ and $||A - A_k||_2 = \sigma_{k+1}$. We may also write $A_k = U\Sigma_k V^T$ where $\Sigma_k = diag(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$.

Proof.

1. **Suppose that A is symmetric, with eigenvalues λ_i and orthonormal eigenvectors u_i. In other words $A = U\Lambda U^T$ is an eigendecomposition of A, with $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$, and $U = [u_1, \ldots, u_n]$, and $UU^T = I$. Then an SVD of A is $A = U\Sigma V^T$, where $\sigma_i = |\lambda_i|$ and $v_i = sign(\lambda_i)u_i$, where $sign(0) = 1$.**

This is true by the definition of the SVD.

2. **The eigenvalues of the symmetric matrix $A^T A$ are σ_i^2. The right singular vectors v_i are corresponding orthonormal eigenvectors.**
$A^T A = V \Sigma U^T U \Sigma V^T = V \Sigma^2 V^T$. This is an eigendecomposition of $A^T A$, with the columns of V the eigenvectors and the diagonal entries of Σ^2 the eigenvalues.

3. **The eigenvalues of the symmetric matrix AA^T are σ_i^2 and $m-n$ zeroes. The left singular vectors u_i are corresponding orthonormal eigenvectors for the eigenvalues σ_i^2. One can take any $m-n$ other orthogonal vectors as eigenvectors for the eigenvalue 0.**
Choose an m-by-$(m-n)$ matrix $\tilde{U}$ so that $[U, \tilde{U}]$ is square and orthogonal. Then write

$$AA^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T = \left[U, \tilde{U}\right] \cdot \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} \cdot \left[U, \tilde{U}\right]^T.$$

This is an eigendecomposition of AA^T.

4. **Let** $H = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$, **where** A **is square and** $A = U\Sigma V^T$ **is the SVD of** A. **Let** $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$, $U = [u_1, \ldots, u_n]$, **and** $V = [\upsilon_1, \ldots, \upsilon_n]$. **Then the 2n eigenvalues of** H **are** $\pm\sigma_i$, **with corresponding unit eigenvectors** $\frac{1}{\sqrt{2}} \begin{bmatrix} \upsilon_i \\ \pm u_i \end{bmatrix}$.

We substitute $A = U\Sigma V^T$ into H to get:

$$H = \begin{bmatrix} 0 & V\Sigma U^T \\ U\Sigma V^T & 0 \end{bmatrix}$$

Choose orthogonal matrix G such that

$$G = \frac{1}{\sqrt{2}} \begin{bmatrix} V & V \\ U & -U \end{bmatrix}$$

It is orthogonal since

$$I = GG^T = \frac{1}{2} \begin{bmatrix} VV^T + VV^T & 0 \\ 0 & UU^T + UU^T \end{bmatrix}$$

Then we observe that

$$G \begin{bmatrix} \Sigma & 0 \\ 0 & \Sigma \end{bmatrix} G^T = \begin{bmatrix} 0 & V\Sigma U^T \\ U\Sigma V^T & 0 \end{bmatrix} = H$$

Then using the spectral theorem we can conclude that the $2n$ eigenvalues of H are $\pm\sigma_i$, with corresponding eigenvectors $\frac{1}{\sqrt{2}} \begin{bmatrix} v_i \\ \pm u_i \end{bmatrix}$.

5. **If A has full rank, the solution of $\min_x \|Ax - b\|_2$ is $x = V\Sigma^{-1}U^T b$.**

$\|Ax - b\|_2^2 = ||U\Sigma V^T x - b||_2^2$. Since A has full rank, so does Σ, and thus Σ is invertible. Now let $[U, \tilde{U}]$ be square and orthogonal as above so

$$\begin{aligned} ||U\Sigma V^T x - b||_2^2 &= \left\| \begin{bmatrix} U^T \\ \tilde{U}^T \end{bmatrix} (U\Sigma V^T x - b) \right\|_2^2 \\ &= \left\| \begin{bmatrix} \Sigma V^T x - U^T b \\ -\tilde{U}^T b \end{bmatrix} \right\|_2^2 \\ &= ||\Sigma V^T x - U^T b||_2^2 + \|\tilde{U}^T b\|_2^2. \end{aligned}$$

This is minimized by making the first term zero, i.e., $x = V\Sigma^{-1}U^T b$.

6. $\|A\|_2 = \sigma_1$. **If A is square and nonsingular, then $\|A^{-1}\|_2^{-1} = \sigma_n$ and $\|A\|_2 \cdot \|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}$.**

It is clear from its definition that the two-norm of a diagonal matrix is the largest absolute entry on its diagonal. Thus, by property of the norm, $\|A\|_2 = \|U^T A V\|_2 = \|U^T U \Sigma V^T V\|_2 = \|\Sigma\|_2 = \sigma_1$ and $\|A^{-1}\|_2 = \|V^T A^{-1} U\|_2 = \|\Sigma^{-1}\|_2 = \sigma_n^{-1}$.

Remark:

$\|A^{-1}\|_2 = \|V^T A^{-1} U\|_2 = \|V^T (U \Sigma V^T)^{-1} U\|_2 = \|\Sigma^{-1}\|_2 = \sigma_n^{-1}$.

7. Write $V = [v_1, v_2, \ldots, v_n]$ and $U = [u_1, u_2, \ldots, u_n]$, so $A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T$ (a sum of rank-1 matrices). Then a matrix of rank $k < n$ closest to A (measured with $||\cdot||_2$) is $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ and $||A - A_k||_2 = \sigma_{k+1}$. We may also write $A_k = U\Sigma_k V^T$ where $\Sigma_k = diag(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$.

7. Write $V = [v_1, v_2, \ldots, v_n]$ and $U = [u_1, u_2, \ldots, u_n]$, so $A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T$ (a sum of rank-1 matrices). Then a matrix of rank $k < n$ closest to A (measured with $||\cdot||_2$) is $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ and $||A - A_k||_2 = \sigma_{k+1}$. We may also write $A_k = U\Sigma_k V^T$ where $\Sigma_k = diag(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$.
A_k has rank k by construction and

$$||A - A_k||_2 = \left\| \sum_{i=1}^{n} \sigma_i u_i v_i^T - \sum_{i=1}^{k} \sigma_i u_i v_i^T \right\|$$

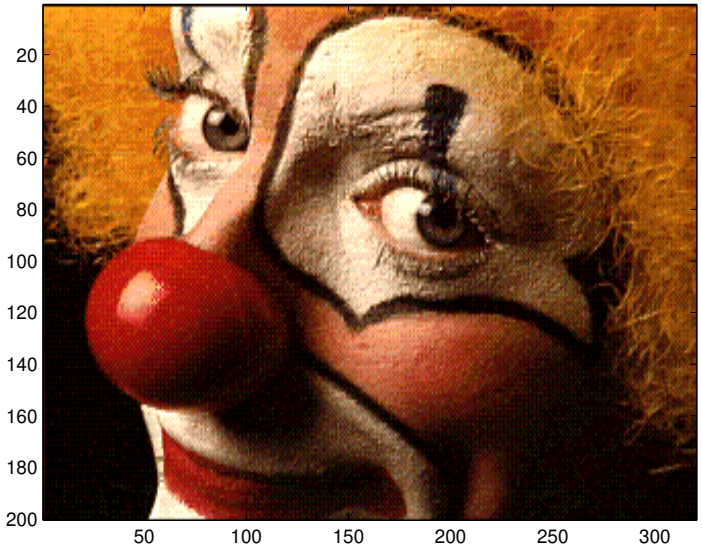
$$= \left\| \sum_{i=k+1}^{n} \sigma_i u_i v_i^T \right\| = \left\| U \begin{bmatrix} 0 & & & \\ & \sigma_{k+1} & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} V^T \right\|_2 = \sigma_{k+1}.$$

It remains to show that there is no closer rank k matrix to A. Let B be any rank k matrix, so its null space has dimension $n-k$. The space spanned by $\{v_1, ..., v_{k+1}\}$ has dimension $k+1$. Since the sum of their dimensions is $(n-k)+(k+1) > n$, these two spaces must overlap. Let h be a unit vector in their intersection. Then

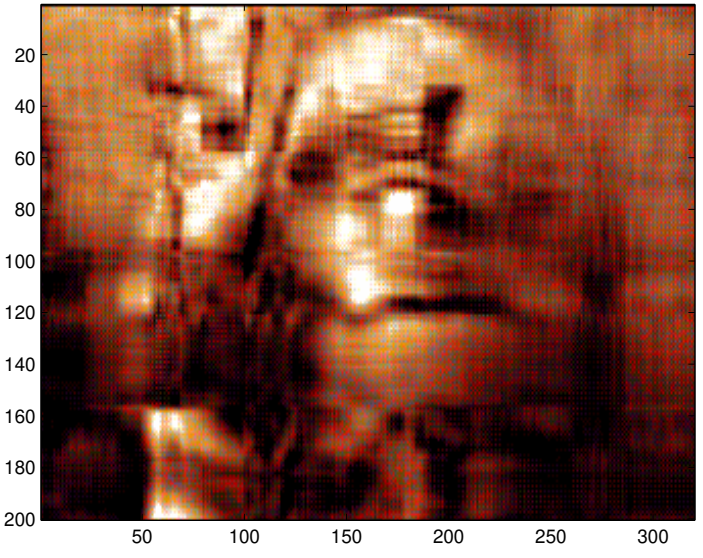
$$\begin{aligned}\|A-B\|_2^2 \geq \|(A-B)h\|_2^2 = \|Ah\|_2^2 = \left\|U\Sigma V^T h\right\|_2^2 \\ = \left\|\Sigma(V^T h)\right\|_2^2 \geq \sigma_{k+1}^2 \left\|V^T h\right\|_2^2 = \sigma_{k+1}^2.\end{aligned}$$

□

Example of application of linear systems: image compression using SVD



a) Original image



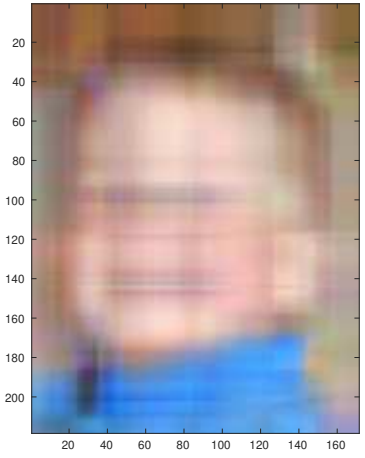
b) Rank k=20 approximation

Example of application of linear systems: image compression using SVD in Matlab

See path for other pictures:
/matlab-2012b/toolbox/matlab/demos
load clown.mat;
Size(X) $= m \times n = 320 \times 200$ pixels.
[U,S,V] = svd(X);
colormap(map);
k=20;
image(U(:,1:k)*S(1:k,1:k)*V(:,1:k)');
Now: size(U)= $m \times k$, size(V)= $n \times k$.

Image compression using SVD in Matlab

a) Original image



b) Rank k=4 approximation

b) Rank k=5 approximation

c) Rank k=6 approximation

d) Rank k=10 approximation

d) Rank k=15 approximation

Example of application of linear systems: image compression using SVD for arbitrary image

To get image on the previous slide, I took picture in jpg-format and loaded it in matlab. You can also try to use following matlab code for your own pictures:

```
A = imread('Child.jpg');
Real size of A: size(A) ans= 218 171 3
figure(1); image(DDA);
DDA=im2double(A);
[U1,S1,V1] = svd(DDA(:,:,1)); [U2,S2,V2] = svd(DDA(:,:,2));
[U3,S3,V3] = svd(DDA(:,:,3));
k=15;
svd1 = U1(:,1:k)*S1(1:k,1:k)*V1(:,1:k)';
svd2 = U2(:,1:k)*S2(1:k,1:k)*V2(:,1:k)';
svd3 = U3(:,1:k)*S3(1:k,1:k)*V3(:,1:k)';
DDAnew = zeros(size(DDA));
DDAnew(:,:,1) = svd1; DDAnew(:,:,2) = svd2; DDAnew(:,:,3) = svd3;
figure(2); image(DDAnew);
```

Perturbation Theory for the Least Squares Problem

When A is not square, we define its condition number with respect to the 2-norm to be $k_2(A) \equiv \sigma_{max}(A)/\sigma_{min}(A)$. This reduces to the usual condition number when A is square. The next theorem justifies this definition.

THEOREM *Suppose that A is m-by-n with $m \geq n$ and has full rank. Suppose that x minimizes $\|Ax - b\|_2$. Let $r = Ax - b$ be the residual. Let $\tilde{x}$ minimize $\|(A + \delta A)\tilde{x} - (b + \delta b)\|_2$. Assume $\epsilon \equiv \max(\frac{\|\delta A\|_2}{\|b\|_2}, \frac{\|\delta b\|_2}{\|b\|_2}) < \frac{1}{k_2(A)} = \frac{\sigma_{min}(A)}{\sigma_{max}(A)}$. Then*

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \epsilon \cdot \left\{ \frac{2 \cdot k_2(A)}{\cos\theta} + \tan\theta \cdot k_2^2(A) \right\} + O(\epsilon^2) \equiv \epsilon \cdot k_{LS} + O(\epsilon^2),$$

where $\sin\theta = \frac{\|r\|_2}{\|b\|_2}$. In other words, θ is the angle between the vectors b and Ax and measures whether the residual norm $\|r\|_2$ is large (near $\|b\|$) or small (near 0). k_{LS} is the condition number for the least squares problem.

Sketch of Proof. Expand $\tilde{x} = ((A + \delta A)^T(A + \delta A))^{-1}(A + \delta A)^T(b + \delta b)$ in powers of δA and δb. Then remove all non-linear terms, leave the linear terms for δA and δb. $\square$

Introduction

- Machine learning is a field of artificial intelligence which gives computer systems the ability to "learn" using available data.
- We will study linear and polynomial classifiers and some artificial neural networks algorithms (multilayer perceptron). We will discover convergence for all these algorithms and compare their performance with respect to applicability, reliability, accuracy, and efficiency. Programs written in Matlab will demonstrate performance for every algorithm.
- Studied algorithms should be applied in comp.ex.3 in numerical studies and comparison of different machine learning algorithms to detect inter-class boundaries.

Reference literature: Miroslav Kurbat, *An Introduction to Machine Learning*, Springer, 2017.

Christopher M. Bishop, *Pattern recognition and machine learning*, Springer, 2009.

L. Beilina, E. Karchevskii, M. Karchevskii, *Numerical Linear Algebra: Theory and Applications*, Springer, 2017 – see link to GitHub with Matlab code

Classification problem

- Suppose that we have data points $(x_i, y_i), i = 1, ..., m$. These points are separated into two classes A and B. Assume that these classes are linearly separable.

  Definition

  Let A and B are two data sets of points in an n-dimensional Euclidean space. Then A and B are linearly separable if there exist $n + 1$ real numbers $\omega_1, ..., \omega_n, l$ such that every point $x \in A$ satisfies $\sum_{i=1}^{n} \omega_i x_i > l$ and every point $x \in B$ satisfies $\sum_{i=1}^{n} \omega_i x_i < -l$.

- Our goal is to find the decision line which will separate these two classes. This line will also predict in which class will the new point fall.

Least squares and classification

Least squares can be used for classification problems appearing in machine learning algorithms.

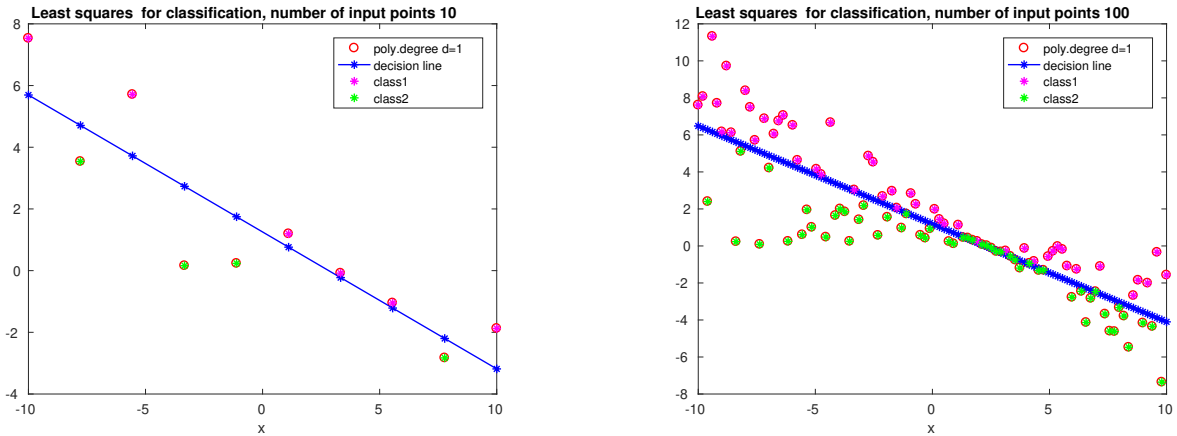


Figure: Examples of working least squares for classification.

Least squares minimization $\min_x \|Ax - b\|_2^2$ for classification is working fine when we know that two classes are linearly separable.

Least squares and classification

Least squares can be used for classification problems appearing in machine learning algorithms.

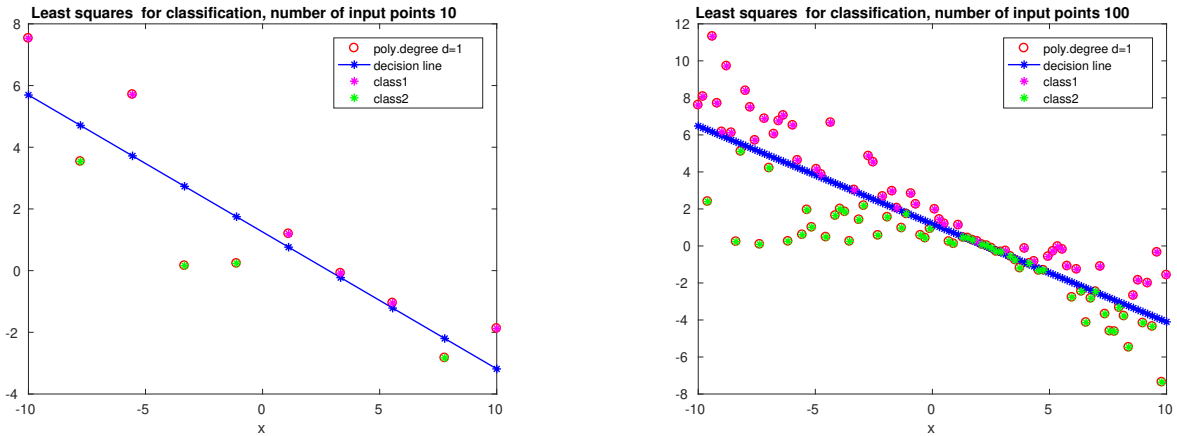


Figure: Examples of working least squares for classification.

Least squares minimization $\min_x \|Ax - b\|_2^2$ for classification is working fine when we know that two classes are linearly separable.

Least squares and classification

Least squares can be used for this classification problem. Let us consider two-class model. Let the first class consisting of l points with coordinates $(x_i, y_i), i = 1, ..., l$ is described by it's linear model

$$f_1(x, c) = c_{1,1}\phi_1(x) + c_{2,1}\phi_2(x) + ... + c_{n,1}\phi_n(x). \quad (1)$$

Let the second class consisting of k points with coordinates $(x_i, y_i), i = 1, ..., k$ is also described by the same linear model

$$f_2(x, c) = c_{1,2}\phi_1(x) + c_{2,2}\phi_2(x) + ... + c_{n,2}\phi_n(x). \quad (2)$$

Here, functions $\phi_j(x), j = 1, ..., n$ are called basis functions. Our goal is to find the vector of parameters $c = c_{i,1} = c_{i,2}, i = 1, ..., n$ of the size n which will fit best to the data $y_i, i = 1, ..., m, m = k + l$ of both model functions, $f_1(x_i, c), i = 1, ..., l$ and $f_2(x_i, c), i = 1, ..., k$ with $f(x, c) = [f_1(x_i, c), f_2(x_i, c)]$ such that

$$\min_c \sum_{i=1}^{m} (y_i - f(x_i, c))^2 \quad (3)$$

with $m = k + l$. If the function $f(x, c)$ is linear then we can solve the problem (3) using least squares method.

Least squares and classification

Let now the matrix A in $Ax = b$ will have entries $a_{ij} = \phi_j(x_i), i = 1, ..., m; j = 1, ..., n$, and vector b will be such that $b_i = y_i, i = 1, ..., m$. Then a linear data fitting problem takes the form

$$Ac = b \tag{4}$$

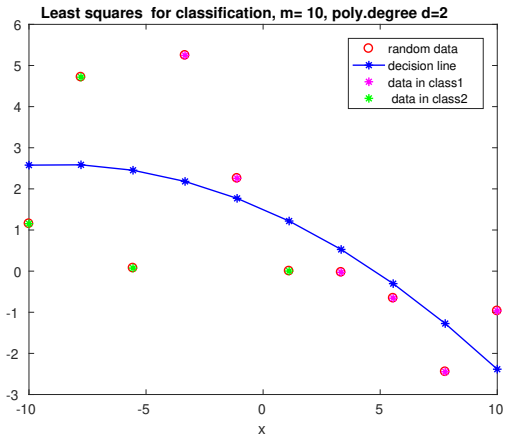
Elements of the matrix A are created by basis functions $\phi_j(x), j = 1, ..., n$. Solution of (4) can be found by the method of normal equations:

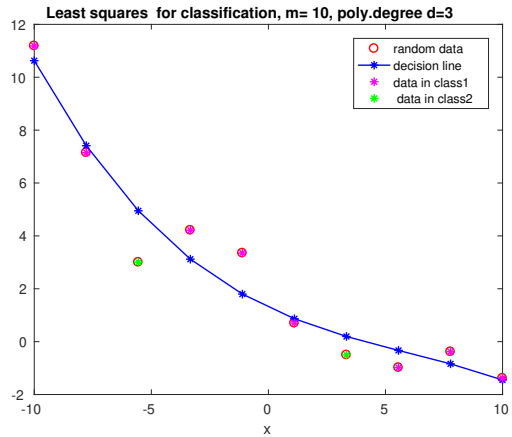
$$c = (A^T A)^{-1} A^T b = A^+ b \tag{5}$$

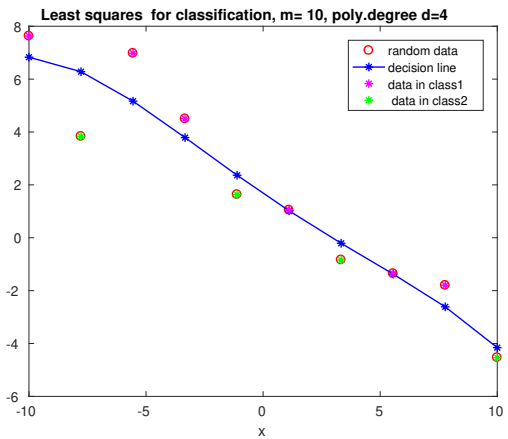
Different basis functions can be chosen. We have considered $\phi_j(x) = x^{j-1}, j = 1, ..., n$ in the problem of fitting to a polynomial. The matrix A constructed by these basis functions is a Vandermonde matrix. Linear splines (or hat functions) and bellsplines also can be used as basis functions.

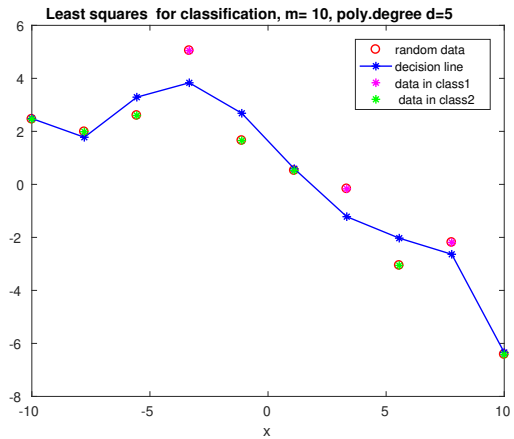
Least squares and classification: example

Least squares minimization for classification is working fine when we know that two classes are linearly separable. Higher degree of polynomial separates two classes better. However, since Vandermonde's matrix can be ill-conditioned for high degrees of polynomial, we should carefully choose appropriate polynomial to fit data.









Least squares and classification

Examples below present computation of decision line for separation of two classes with $m = 100$ using basis functions $\phi_j(x) = x^{j-1}, j = 1, ..., d$, where d is degree of the polynomial.

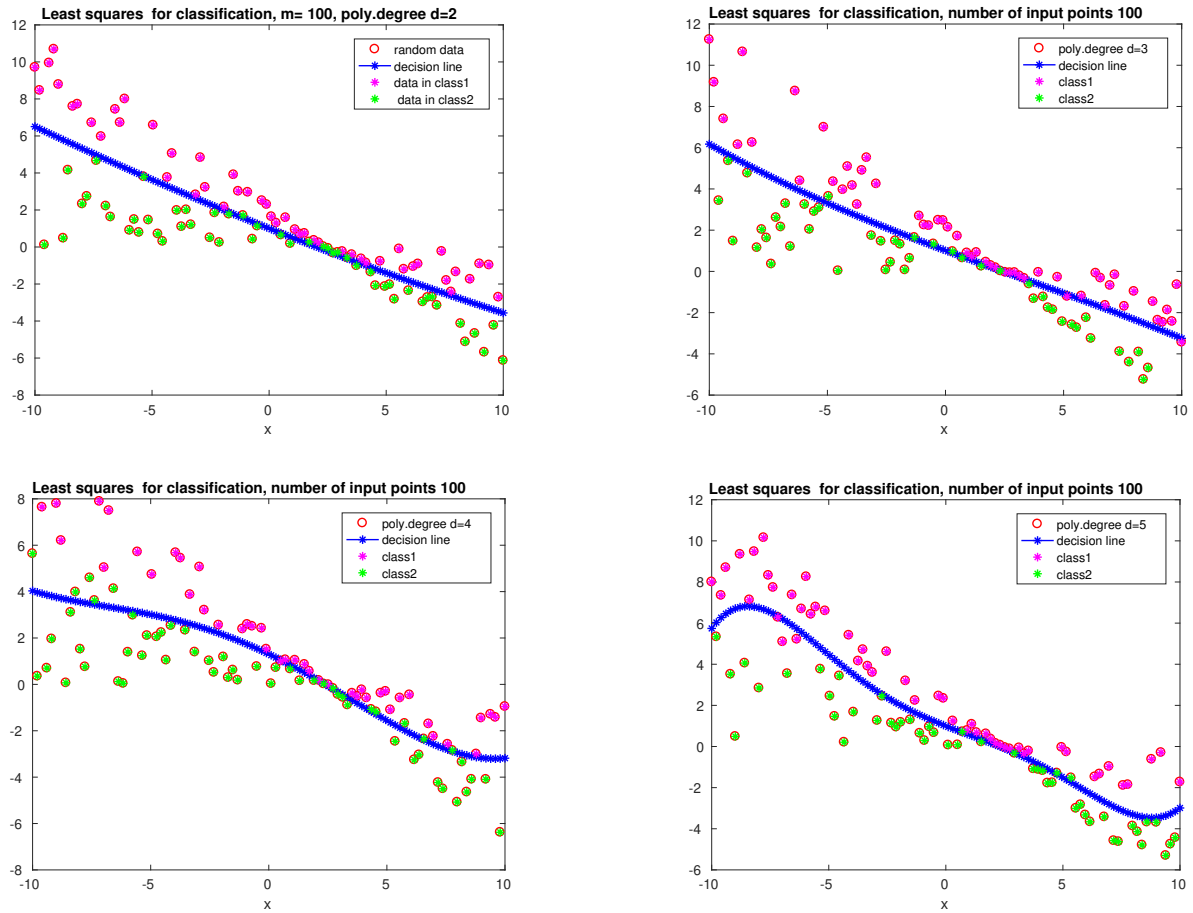


Figure: Examples of working least squares for separation of two classes.

Machine learning algorithms: linear and polynomial classifiers

Let us consider boolean domains where each attribute is true or false. and we will represent *true* by 1 and *false* by 0.
Below we present a table where is presented a boolean domain with two classes and two boolean attributes (here, true is 1 and false is 0).

x	y	Class
1	1	positive
1	0	negative
0	1	negative
0	0	negative

We observe that these two classes can be separated by linear equation

$$\omega_1 + \omega_2 x + \omega_3 y = 0. \tag{6}$$

Our goal is to find weights $\omega_1, \omega_2, \omega_3$ in order to determine the decision line $y(x)$. The decision line which will separate two classes will have the form $y(x) = (-\omega_1 - \omega_2 x)/\omega_3$.

Linear and polynomial classifiers

On the figure below, two classes should be separated:one example we labeled as positive class, another one as negative. In this case, two classes can be separated by linear equation

$$\omega_1 + \omega_2 x + \omega_3 y = 0 \tag{7}$$

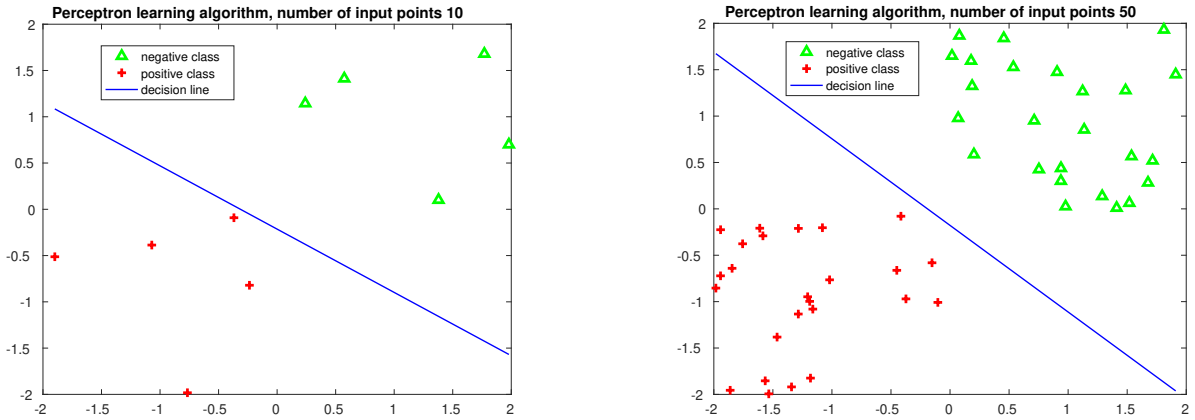


Figure: Examples of working perceptron learning algorithm which computes weights for decision line to separate two classes.

Linear and polynomial classifiers

In common case, two classes can be separated by the general equation

$$\omega_0 + \omega_1 x_1 + \omega_2 x_2 + ... + \omega_n x_n = 0 \tag{8}$$

which also can be written as

$$\sum_{i=0}^{n} \omega_i x_i = 0 \tag{9}$$

with $x_0 = 1$. If $n = 2$ then the above equation defines a line, if $n = 3$ - plane, if $n > 3$ - hyperplane. Our problem is to determine weights ω_i and the task of machine learning is to determine their appropriate values. Weights $\omega_i, i = 1, ..., n$ determine the angle of the hyperplane, ω_0 is called bias and determines the offset, or the hyperplanes distance from the origin of the system of coordinates.

Machine learning: Perceptron learning algorithm

- Let us assume that every training example $\mathbf{x} = (x_1, ..., x_n)$ is described by n attributes with values $x_i = 0$ or $x_i = 1$.
- We will label positive examples with $c(\mathbf{x}) = 1$ and negative with $c(\mathbf{x}) = 0$.
- Let us denote by $h(\mathbf{x})$ the classifier's hypothesis which also will have binary values $h(\mathbf{x}) = 1$ or $h(\mathbf{x}) = 0$.
- We will also assume that all examples where $c(\mathbf{x}) = 1$ are linearly separable from examples where $c(\mathbf{x}) = 0$.

Machine learning: Perceptron learning algorithm

- Step 0. Initialize weights ω_i to small random numbers.
- Step 1. If $\sum_{i=0}^{n} \omega_i x_i > 0$ we will say that the example is positive and $h(x) = 1$.
- Step 2. If $\sum_{i=0}^{n} \omega_i x_i < 0$ we will say the the example is negative and $h(x) = 0$.
- Step 3. Update every weight using the formula

$$\omega_i = \omega_i + \eta \cdot [c(\mathbf{x}) - h(\mathbf{x})] \cdot x_i.$$

- Step 4. If $c(\mathbf{x}) = h(\mathbf{x})$ for all learning examples - stop. Otherwise return to step 1.

Here, $\eta \in (0, 1]$ is called the learning rate.

Polynomial of the second order

Coefficients of polynomials of the second order can be obtained by the same technique as coefficients for linear classifiers.
The second order polynomial function is:

$$\omega_0 + \omega_1 \underbrace{x_1}_{z_1} + \omega_2 \underbrace{x_2}_{z_2} + \omega_3 \underbrace{x_1^2}_{z_3} + \omega_4 \underbrace{x_1 x_2}_{z_4} + \omega_5 \underbrace{x_2^2}_{z_5} = 0 \qquad (10)$$

This polynomial can be converted to the linear classifier if we introduce notations:

$$z_1 = x_1, z_2 = x_2, z_3 = x_1^2, z_4 = x_1 x_2, z_5 = x_2^2.$$

Then equation (10) can be written in new variables as

$$\omega_0 + \omega_1 z_1 + \omega_2 z_2 + \omega_3 z_3 + \omega_4 z_4 + \omega_5 z_5 = 0 \qquad (11)$$

which is already linear function. Thus, the Perceptron learning algorithm can be used to determine weights $\omega_0, ..., \omega_5$ in (11).

Polynomial of the second order

Suppose that you have determined weights $\omega_0, ..., \omega_5$ in (11). To present the decision line you need to solve the quadratic equation for x_2:

$$\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_1^2 + \omega_4 x_1 x_2 + \omega_5 x_2^2 = 0 \tag{12}$$

with known weights $\omega_0, ..., \omega_5$ and known x_1. We can rewrite (12) as

$$\underbrace{\omega_5}_{a} x_2^2 + x_2 \underbrace{(\omega_2 + \omega_4 x_1)}_{b} + \underbrace{\omega_0 + \omega_1 x_1}_{c} = 0 \tag{13}$$

or as

$$a x_2^2 + b x_2 + c = 0. \tag{14}$$

Solutions of (14) will be

$$x_2 = \frac{-b \pm \sqrt{D}}{2a},$$
$$D = b^2 - 4ac.$$

Perceptron learning algorithm for polynomial of the second order: example

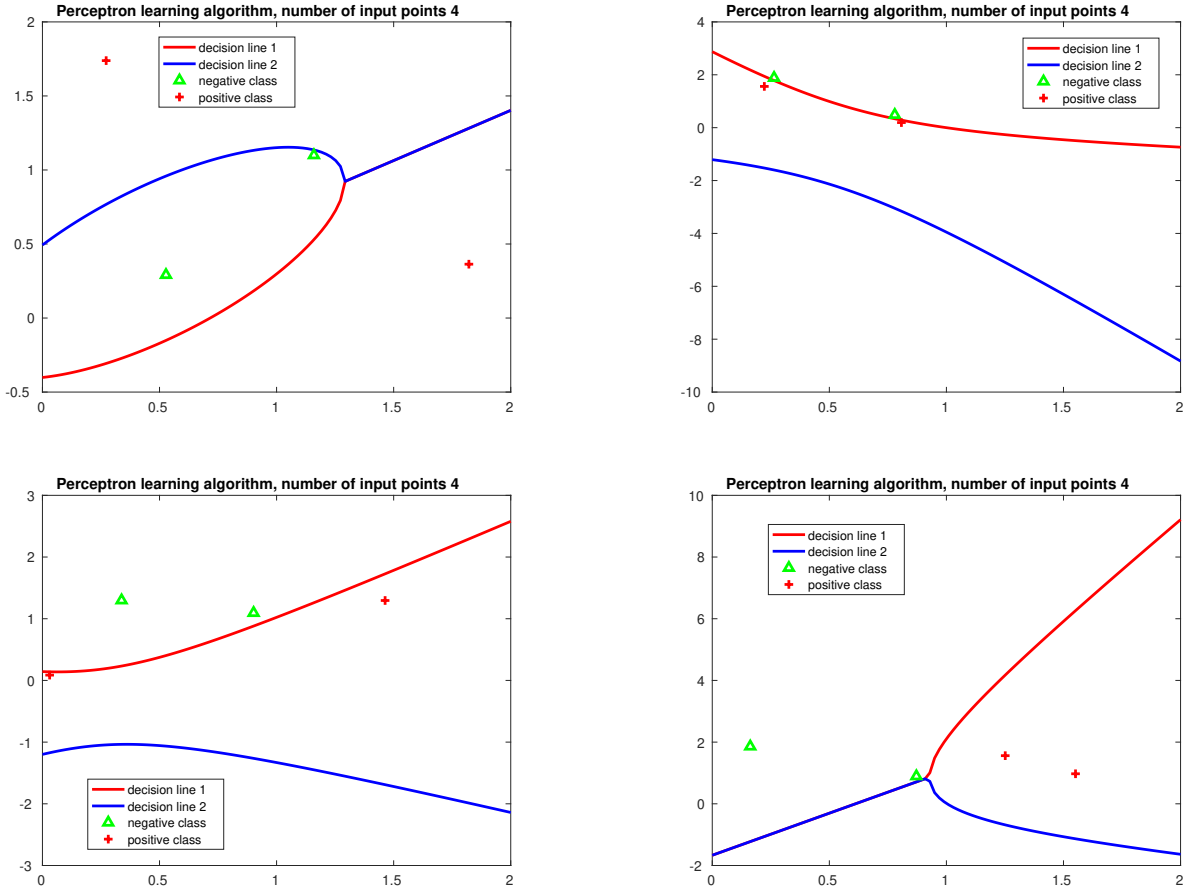


Figure: Separation of two classes by polynomials of the second order for 4 points.

Perceptron learning algorithm for polynomial of the second order: example

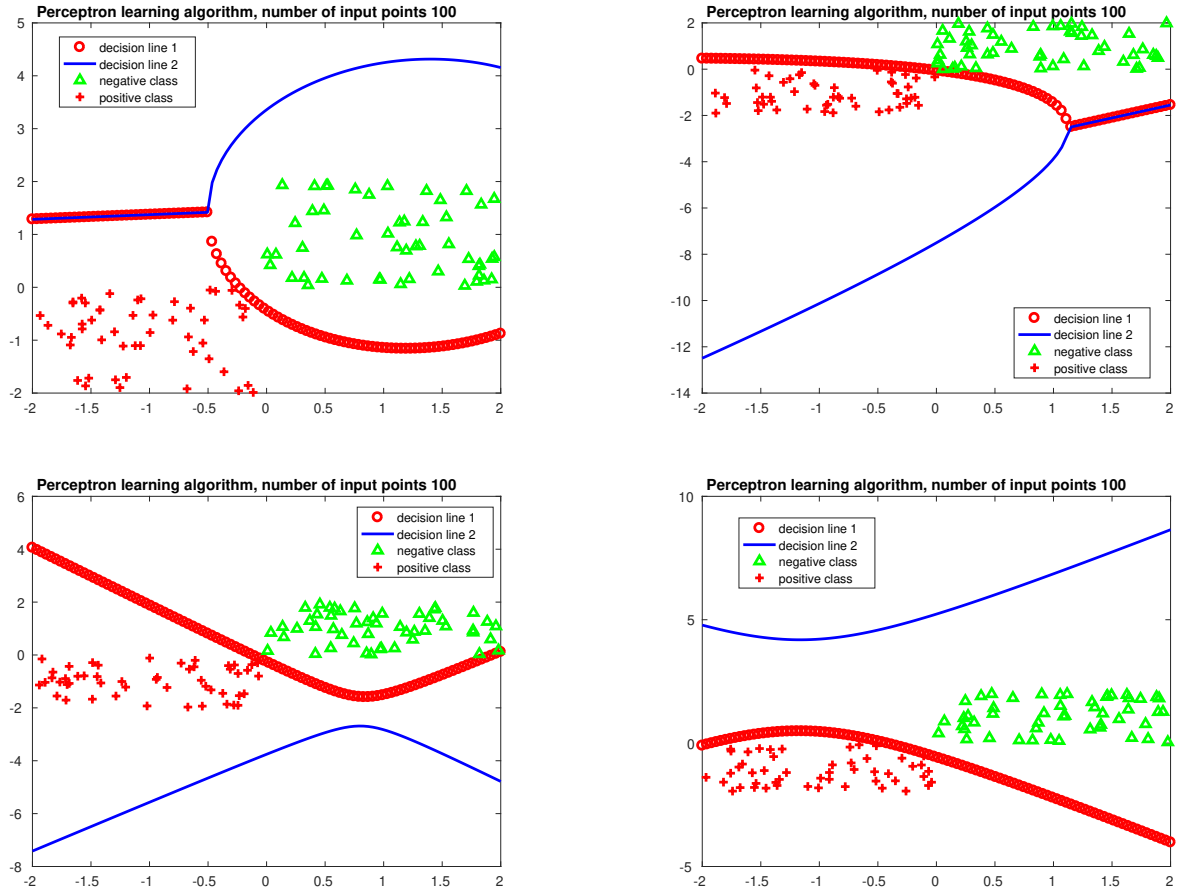


Figure: Separation of two linearly separated classes by polynomials of the second order for 100 points.

Perceptron learning algorithm for polynomial of the second order: example

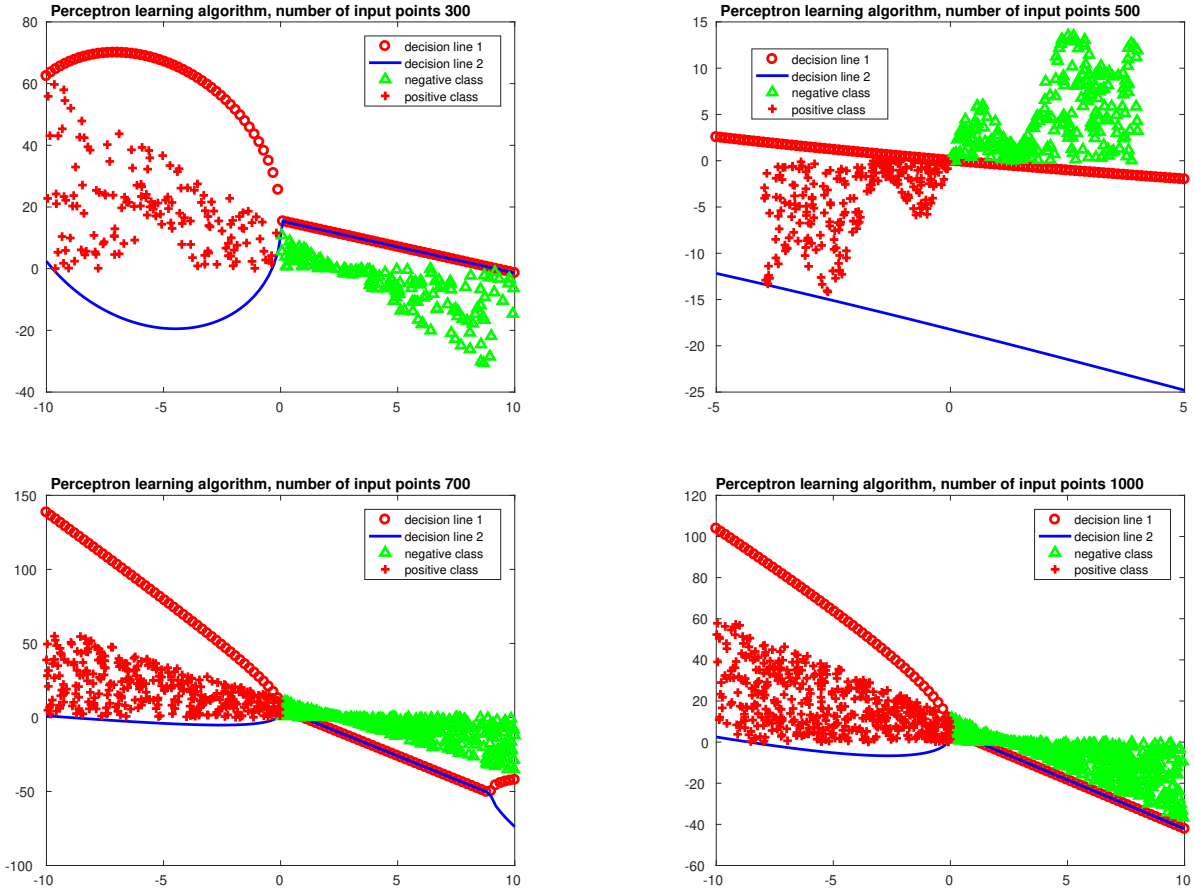


Figure: Separation of two linearly separated classes by polynomials of the second order.

Perceptron learning algorithm for polynomial of the second order: example

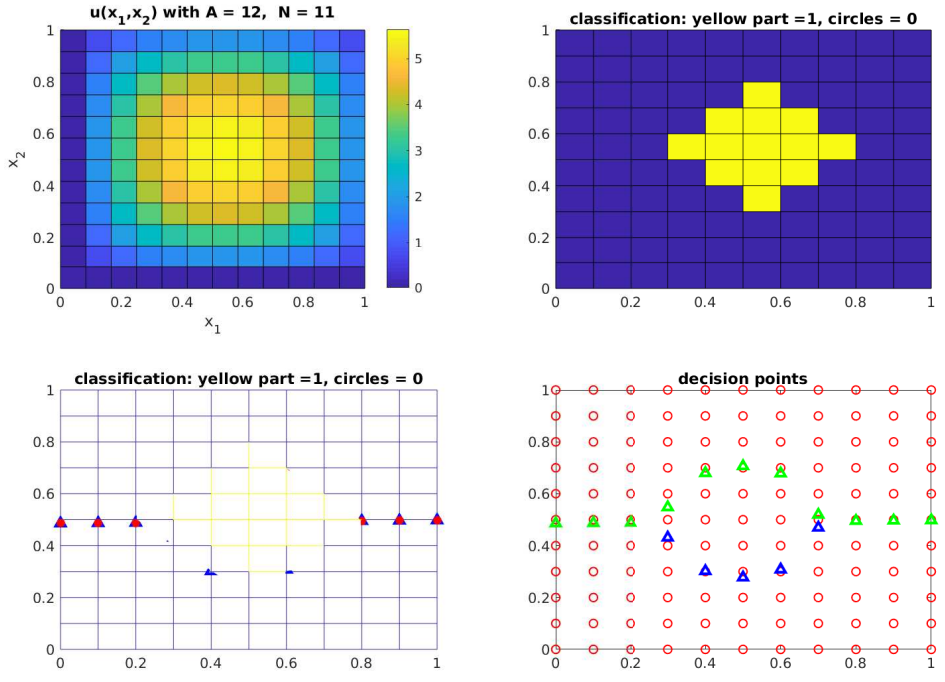


Figure: Separation of the solution of Poisson equation in 2D on the square (see example 8.1.3 in the course book) by polynomials of the second order.

Perceptron learning algorithm for polynomial of the second order: example

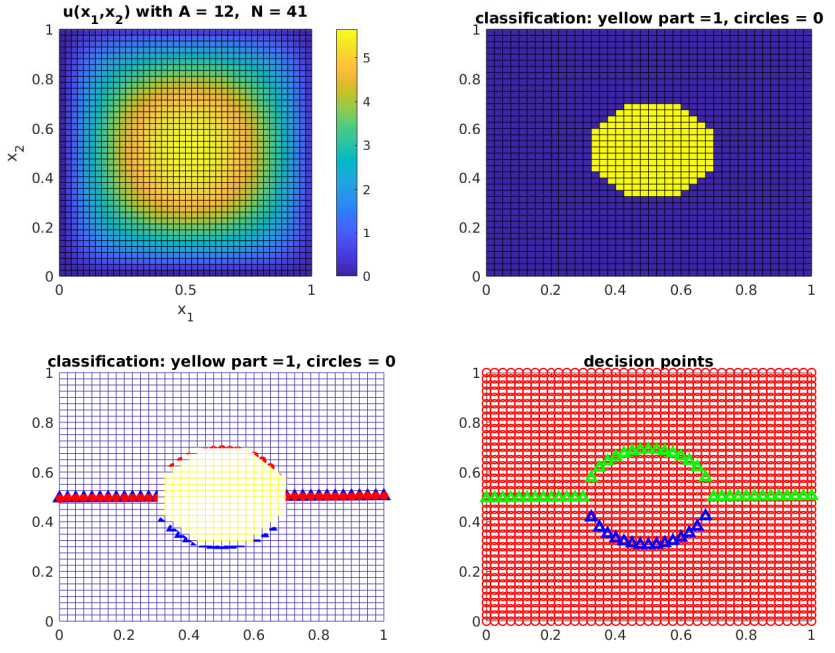


Figure: Separation of the solution of Poisson equation in 2D on the square (see example 8.1.3 in the course book) by polynomials of the second order.

Machine learning: WINNOW learning algorithm

Perceptron learning algorithm used additive rule, while WINNOW algorithm uses multiplicative rule: weights are multiplied in this rule. We will again assume that all examples where $c(\mathbf{x}) = 1$ are linearly separable from examples where $c(\mathbf{x}) = 0$. Main steps in the WINNOW learning algorithm are:
Step 0. Initialize weights $\omega_i = 1$. Choose parameter $\alpha > 1$, usually $\alpha = 2$.
Step 1. If $\sum_{i=0}^{n} \omega_i x_i > 0$ we will say that the example is positive and $h(x) = 1$.
Step 2. If $\sum_{i=0}^{n} \omega_i x_i < 0$ we will say the the example is negative and $h(x) = 0$.
Step 3. Update every weight using the formula

$$\omega_i = \omega_i * \alpha^{c(\mathbf{x}) - h(\mathbf{x})}.$$

Step 4. If $c(\mathbf{x}) = h(\mathbf{x})$ for all learning examples - stop. Otherwise return to step 1.

Perceptron learning algorithm vs. WINNOW: example

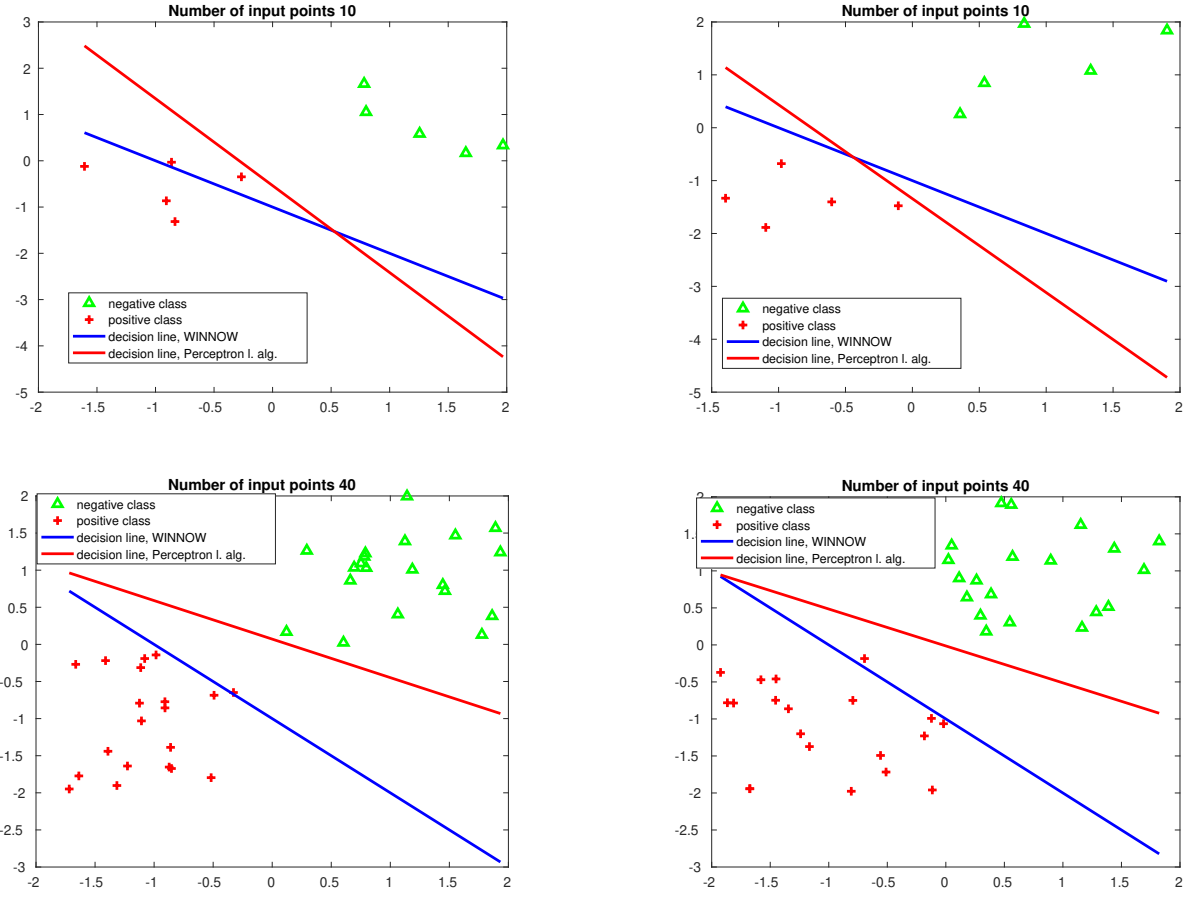


Figure: Comparison of two classification algorithms for separation of two classes: Perceptron learning algorithm (red line) and WINNOW (blue line).

Artificial neural networks

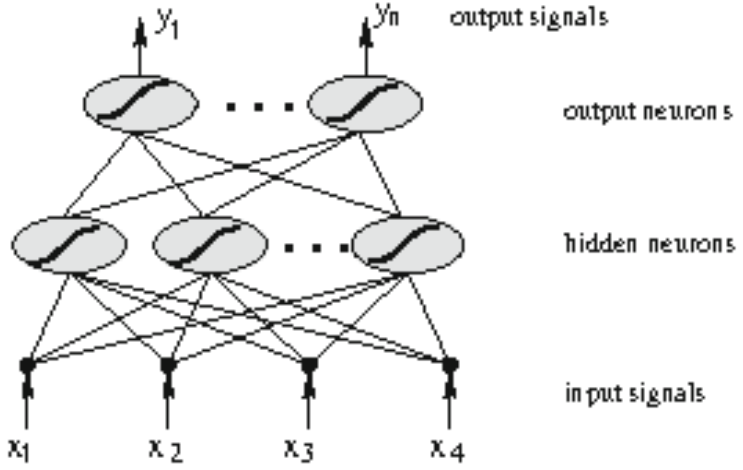


Figure: Example of neural network which contains two interconnected layers (M. Kurbat, *An Introduction to machine learning*, Springer, 2017.)

- In an artificial neural network simple units - neurons- are interconnected by weighted links into into structures of high perfomance.
- We will consider multilayer perceptrons and radial basis function networks.

Neurons

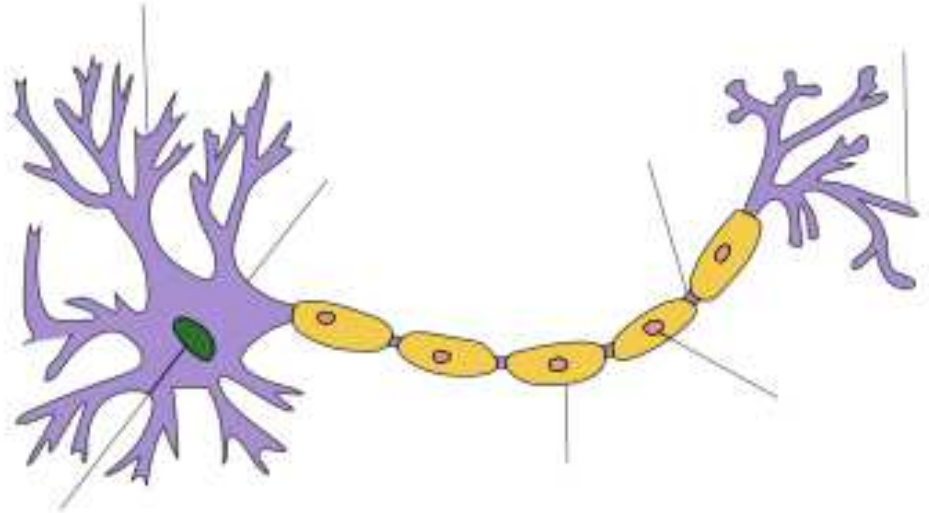


Figure: Structure of a typical neuron (Wikipedia).

- A neuron, also known as a nerve cell, is an electrically excitable cell that receives, processes, and transmits information through electrical and chemical signals. These signals between neurons occur via specialized connections called synapses.
- An artificial neuron is a mathematical function which presents a model of biological neurons, a neural network.

Artificial neurons

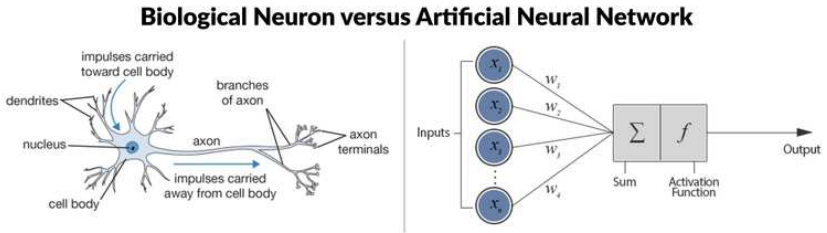


Figure: Perceptron neural network consisting of one neuron (source: DataCamp(datacamp.com)).

- Artificial neurons are elementary units in an artificial neural network. The artificial neuron receives one or more inputs and sums them to produce an output (or activation, representing a neuron's action potential which is transmitted along its axon).
- Each input is separately weighted by weights ω_{kj}, and the sum $\sum_k \omega_{kj} x_k$ is passed as an argument $\Sigma = \sum_k \omega_{kj} x_k$ through a non-linear function $f(\Sigma)$ which is called the activation function or transfer function.
- Assume that attributes x_k are normalized and belong to the interval $[-1, 1]$.

Artificial neurons: transfer functions

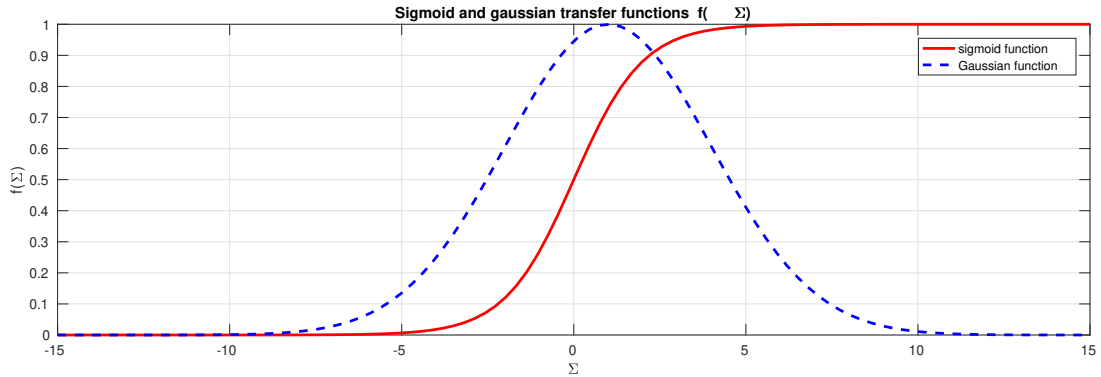


Figure: Sigmoid and Gaussian (for $b = 1, \sigma = 3$ in (16)) transfer functions.

- Different transfer (or activation) functions $f(\Sigma)$ with $\Sigma = \sum_k \omega_{kj} x_k$ are used. We will study sigmoid and gaussian functions.
- Sigmoid function:

$$f(\Sigma) = \frac{1}{1 + e^{-\Sigma}} \tag{15}$$

- Gaussian function centered at b for a given variance σ^2

$$f(\Sigma) = \frac{e^{-(\Sigma - b)^2}}{2\sigma^2} \tag{16}$$

Forward propagation

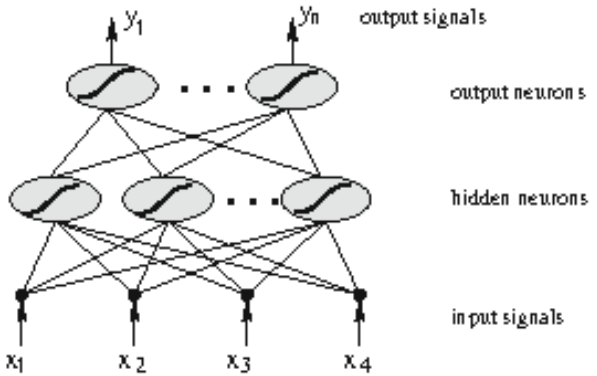


Figure: Example of neural network called multilayer perceptron (one hidden layer of neurons and one output layer). (M. Kurbat, *An Introduction to machine learning*, Springer, 2017.)

- Neurons in adjacent layer are fully interconnected.
- Forward propagation is implemented as

$$y_i = f(\Sigma_j \omega_{ji}^{(1)} x_j) = f(\Sigma_j \omega_{ji}^{(1)} \underbrace{f(\Sigma_k \omega_{kj}^{(2)} x_k)}_{x_j}), \qquad (17)$$

where $\omega_{ji}^{(1)}$ and $\omega_{kj}^{(2)}$ are weights of the output and the hidden neurons, respectively, f is the transfer function.

Example of forward propagation through the network

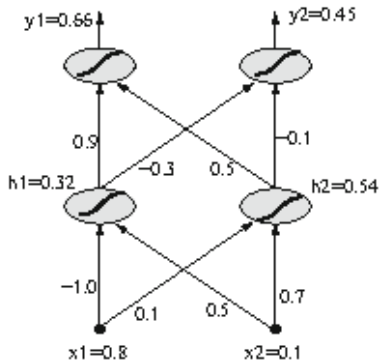


Figure: Source: M. Kurbat, *An Introduction to machine learning*, Springer, 2017.

- Using inputs x_1, x_2 compute inputs of hidden-layer neurons:

$$x_1^{(2)} = 0.8 * (-1.0) + 0.1 * 0.5 = -0.75, \;\; x_2^{(2)} = 0.8 * 0.1 + 0.1 * 0.7 = 0.15$$

- Compute transfer function (sigmoid $f(\Sigma) = \frac{1}{1+e^{-\Sigma}}$ in our case):

$$h_1 = f(x_1^{(2)}) = 0.32, \;\; h_2 = f(x_2^{(2)}) = 0.54.$$

- Compute input of output-layer neurons

$$x_1^{(1)} = 0.32 * 0.9 + 0.54 * 0.5 = 0.56, \;\; x_2^{(1)} = 0.32 * (-0.3) + 0.54 * (-0.1) = -0.15.$$

- Compute outputs of output-layer neurons using transfer function (sigmoid in our case):

$$y_1 = f(x_1^{(1)}) = 0.66, \;\; y_2 = f(x_2^{(1)}) = 0.45.$$

Backpropagation of error through the nework

Our goal is to find optimal weights $\omega_{ji}^{(1)}$ and $\omega_{kj}^{(2)}$ in forward propagation

$$y_i = f(\Sigma_j \omega_{ji}^{(1)} x_j) = f(\Sigma_j \omega_{ji}^{(1)} \underbrace{f(\Sigma_k \omega_{kj}^{(2)} x_k)}_{x_j}). \tag{18}$$

To do this we introduce functional

$$F(\omega_{ji}^{(1)}, \omega_{kj}^{(2)}) = \frac{1}{2}\|t_i - y_i\|^2 = \frac{1}{2}\sum_{i=1}^{m}(t_i - y_i)^2. \tag{19}$$

Here, $t = t(x)$ is the target vector which depends on the concrete example x. In the domain with m classes the target vector $t = (t_1(x), ..., t_m(x))$ consists of m binary numbers such that

$$t_i(x) = \begin{cases} 1, & \text{example } x \text{ belongs to } i\text{-th class,} \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Examples of target vector and mean square error

Let there exist three different classes c_1, c_2, c_3. Let the example x belongs to the class c_2. Then the target vector is $t = (t_1, t_2, t_3) = (0, 1, 0)$.
The mean square error is defined as

$$E = \frac{1}{m}\|t_i - y_i\|^2 = \frac{1}{m}\sum_{i=1}^{m}(t_i - y_i)^2. \tag{21}$$

Let us assume that we have two different networks to choose from, every network with 3 output neurons corresponding to classes c_1, c_2, c_3. Let $t = (t_1, t_2, t_3) = (0, 1, 0)$ and for the example x the first network output is $y_1 = (0.5, 0.2, 0.9)$ and the second network output is $y_2 = (0.6, 0.6, 0.7)$.

$$E_1 = \frac{1}{3}\sum_{i=1}^{3}(t_i - y_i)^2 = \frac{1}{3}((0 - 0.5)^2 + (1 - 0.2)^2 + (0 - 0.9)^2)) = 0.57,$$

$$E_2 = \frac{1}{3}\sum_{i=1}^{3}(t_i - y_i)^2 = \frac{1}{3}((0 - 0.6)^2 + (1 - 0.6)^2 + (0 - 0.7)^2)) = 0.34.$$

Since $E_2 < E_1$ then the second network is less wrong on the example x than the first network.

Backpropagation of error through the nework

To find minimum of the functional (19) $F(\omega)$ with $\omega = (\omega_{ji}^{(1)}, \omega_{kj}^{(2)})$, recall it below:

$$F(\omega) = F(\omega_{ji}^{(1)}, \omega_{kj}^{(2)}) = \frac{1}{2}\|t_i - y_i\|^2 = \frac{1}{2}\sum_{i=1}^{m}(t_i - y_i)^2, \tag{22}$$

we need to solve the minimization problem

$$\min_{\omega} F(\omega) \tag{23}$$

and find a stationary point of (22)) with respect to ω such that

$$F'(\omega)(\bar{\omega}) = 0, \tag{24}$$

where $F'(\omega)$ is the Fréchet derivative such that

$$F'(\omega)(\bar{\omega}) = F'_{\omega_{ji}^{(1)}}(\omega)(\bar{\omega}_{ji}^{(1)}) + F'_{\omega_{kj}^{(2)}}(\omega)(\bar{\omega}_{kj}^{(2)}). \tag{25}$$

Backpropagation of error through the nework

Recall now that y_i in the functional (22) is defined as

$$y_i = f(\sum_j \omega_{ji}^{(1)} x_j) = f(\sum_j \omega_{ji}^{(1)} \underbrace{f(\sum_k \omega_{kj}^{(2)} x_k)}_{x_j}). \tag{26}$$

Thus, if the transfer function f in (28) is sigmoid, then

$$\begin{aligned} F'_{\omega_{ji}^{(1)}}(\omega)(\bar{\omega}_{ji}^{(1)}) &= (t_i - y_i) \cdot y'_{i\,(\omega_{ji}^{(1)})}(\bar{\omega}_{ji}^{(1)}) \\ &= (t_i - y_i) \cdot x_j \cdot f(\sum_j \omega_{ji}^{(1)} x_j)(1 - f(\sum_j \omega_{ji}^{(1)} x_j))(\bar{\omega}_{ji}^{(1)}) \\ &= (t_i - y_i) \cdot x_j \cdot y_i(1 - y_i))(\bar{\omega}_{ji}^{(1)}), \end{aligned} \tag{27}$$

since for the sigmoid function $f'(\Sigma) = f(\Sigma)(1 - f(\Sigma))$ (prove this).

Backpropagation of error through the nework

Again, since

$$y_i = f(\sum_j \omega_{ji}^{(1)} x_j) = f(\sum_j \omega_{ji}^{(1)} \underbrace{f(\sum_k \omega_{kj}^{(2)} x_k)}_{x_j}). \tag{28}$$

for the sigmoid transfer function f we also get

$$\begin{aligned} F'_{\omega_{kj}^{(2)}}(\omega)(\bar{\omega}_{kj}^{(2)}) &= (t_i - y_i) \cdot y'_{i\,(\omega_{kj}^{(2)})}(\bar{\omega}_{kj}^{(2)}) \\ &= \left[\underbrace{h_j(1-h_j)}_{f'(h_j)} \cdot \left[\sum_i \underbrace{y_i(1-y_i)}_{f'(y_i)} (t_i - y_i) \omega_{ji}^{(1)} \right] \cdot x_k \right] (\bar{\omega}_{kj}^{(2)}), \end{aligned} \tag{29}$$

since for the sigmoid function f we have:
$f'(h_j) = f(h_j)(1 - f(h_j)), f'(y_i) = f(y_i)(1 - f(y_i))$ (prove this). Hint:
$h_j = f(\sum_k \omega_{kj}^{(2)} x_k), y_i = f(\sum_j \omega_{ji}^{(1)} x_j)$.

Backpropagation of error through the nework

Usually, $F'_{\omega_{ji}^{(1)}}(\omega)/x_j, F'_{\omega_{kj}^{(2)}}(\omega)/x_k$ in (27), (29) are called responsibilities of output layer neurons and hidden-layer neurons $\delta_i^{(1)}, \delta_i^{(2)}$, respectively, and they are defined as

$$\begin{aligned} \delta_i^{(1)} &= (t_i - y_i) y_i (1 - y_i), \\ \delta_j^{(2)} &= h_j (1 - h_j) \cdot \sum_i \delta_i^{(1)} \omega_{ji}^{(1)}. \end{aligned} \tag{30}$$

By knowingf responsibilities (30), weights can be updates using usual gradient update formulas:

$$\begin{aligned} \omega_{ji}^{(1)} &= \omega_{ji}^{(1)} + \eta \delta_i^{(1)} x_j, \\ \omega_{kj}^{(2)} &= \omega_{kj}^{(2)} + \eta \delta_j^{(2)} x_k. \end{aligned} \tag{31}$$

Here, η is the step size in the gradient update of weights and we use value of learning rate for it such that $\eta \in (0, 1)$.

Algorithm of backpropagation of error through the nework with one hidden layer

- Step 0. Initialize weights and take $l = 1$.
- Step 1. Take example x^l in the input layer and perform forward propagation.
- Step 2. Let $y^l = (y_1^l, \ldots, y_m^l)$ be the output layer and let $t^l = (t_1^l, \ldots, t_m^l)$ be the target vector.
- Step 3. For every output neuron y_i^l, $i = 1, \ldots, m$ calculate its responsibility $(\delta_i^{(1)})^l$ as

$$(\delta_i^{(1)})^l = (t_i^l - y_i^l) y_i^l (1 - y_i^l). \tag{32}$$

- Step 4. For every hidden neuron compute responsibility $(\delta_j^{(2)})^l$ for the network's error as

$$(\delta_j^{(2)})^l = h_j^l (1 - h_j^l) \cdot \sum_i (\delta_i^{(1)})^l (\omega_{ji}^{(1)})^l, \tag{33}$$

where $(\delta_i^{(1)})^l$ are computed using (32).

- Step 5. Update weights with learning rate $\eta^l \in (0, 1)$ as

$$\begin{aligned} (\omega_{ji}^{(1)})^{l+1} &= (\omega_{ji}^{(1)})^l + \eta^l (\delta_i^{(1)})^l x_j^l, \\ (\omega_{kj}^{(2)})^{l+1} &= (\omega_{kj}^{(2)})^l + \eta^l (\delta_j^{(2)})^l x_k^l. \end{aligned} \tag{34}$$

- Step 6. If the mean square error less than tolerance, or $\|(\omega_{ji}^{(1)})^{l+1} - (\omega_{ji}^{(1)})^l\| < \epsilon_1$ and $\|(\omega_{kj}^{(2)})^{l+1} - (\omega_{kj}^{(2)})^l\| < \epsilon_2$ stop, otherwise increase number of iterations of the algorithm $l = l + 1$ and go to the step 1. Here, ϵ_1, ϵ_2 are tolerances chosen by the user.

Example of backpropagation of error through the nework

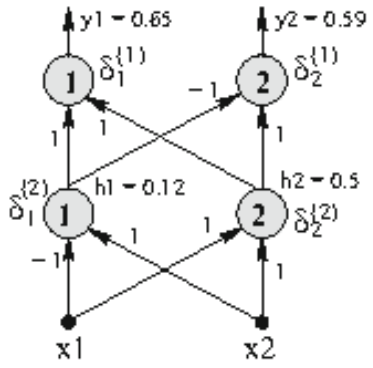


Figure: Source: M. Kurbat, *An Introduction to machine learning*, Springer, 2017.

- Assume that after forward propagation with sigmoid transfer function we have

$$h_1 = f(x_1^{(2)}) = 0.12, \;\; h_2 = f(x_2^{(2)}) = 0.5,$$
$$y_1 = f(x_1^{(1)}) = 0.65, \;\; y_2 = f(x_2^{(1)}) = 0.59.$$

- Let the target vector be $t(x) = (1, 0)$ for the output vector $y = (0.65, 0.59)$.
- Compute responsibility for the output neurons:

$$\sigma_1^{(1)} = y_1 * (1 - y_1)(t_1 - y_1) = 0.65(1 - 0.65)(1 - 0.65) = 0.0796,$$
$$\sigma_2^{(1)} = y_2 * (1 - y_2)(t_2 - y_2) = 0.59(1 - 0.59)(0 - 0.59) = -0.1427$$

Example of backpropagation of error through the nework

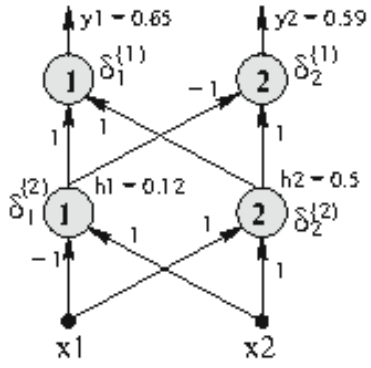


Figure: Source: M. Kurbat, *An Introduction to machine learning*, Springer, 2017.

- Compute the weighted sum for every hidden neuron

$$\delta_1 = \sigma_1^{(1)} w_{11}^{(1)} + \sigma_2^{(1)} w_{12}^{(1)} = 0.0796 * 1 + (-0.1427) * (-1) = 0.2223,$$
$$\delta_2 = \sigma_1^{(1)} w_{21}^{(1)} + \sigma_2^{(1)} w_{22}^{(1)} = 0.0796 * 1 + (-0.1427) * 1 = -0.0631.$$

- Compute responsibility for the hidden neurons for above computed δ_1, δ_2:

$$\sigma_1^{(2)} = h_1(1 - h_1)\delta_1 = -0.0235, \ \sigma_2^{(2)} = h_2(1 - h_2)\delta_2 = 0.0158.$$

Example of backpropagation of error through the nework

- Compute new weights $\omega_{ji}^{(1)}$ for output layer with learning rate $\eta = 0.1$ as:

$$\omega_{11}^{(1)} = \omega_{11}^{(1)} + \eta\sigma_1^{(1)} h_1 = 1 + 0.1 * 0.0796 * 0.12 = 1.00096,$$
$$\omega_{21}^{(1)} = \omega_{21}^{(1)} + \eta\sigma_1^{(1)} h_2 = 1 + 0.1 * 0.0796 * 0.5 = 1.00398,$$
$$\omega_{12}^{(1)} = \omega_{12}^{(1)} + \eta\sigma_2^{(1)} h_1 = -1 + 0.1 * (-0.1427) * 0.12 = -1.0017,$$
$$\omega_{22}^{(1)} = \omega_{22}^{(1)} + \eta\sigma_2^{(1)} h_2 = 1 + 0.1 * (-0.1427) * 0.5 = 0.9929.$$

- Compute new weights $\omega_{kj}^{(2)}$ for hidden layer with learning rate $\eta = 0.1$ as:

$$\omega_{11}^{(2)} = \omega_{11}^{(2)} + \eta\sigma_1^{(2)} x_1 = -1 + 0.1 * (-0.0235) * 1 = -1.0024,$$
$$\omega_{21}^{(2)} = \omega_{21}^{(2)} + \eta\sigma_1^{(2)} x_2 = 1 + 0.1 * (-0.0235) * 1 = 1.0024,$$
$$\omega_{12}^{(2)} = \omega_{12}^{(2)} + \eta\sigma_2^{(2)} x_1 = 1 + 0.1 * 0.0158 * 1 = 1.0016,$$
$$\omega_{22}^{(2)} = \omega_{22}^{(2)} + \eta\sigma_2^{(2)} x_2 = 1 + 0.1 * 0.0158 * (-1) = 0.9984.$$

- Using computed weights for hidden and output layers, one can test a neural network for a new example.

Computer exercise 3 (2 p.)

In this exercise we will study different linear and quadratic classifiers: least squares classifier, perceptron learning algorithm, WINNOW algorithm using training sets described below.

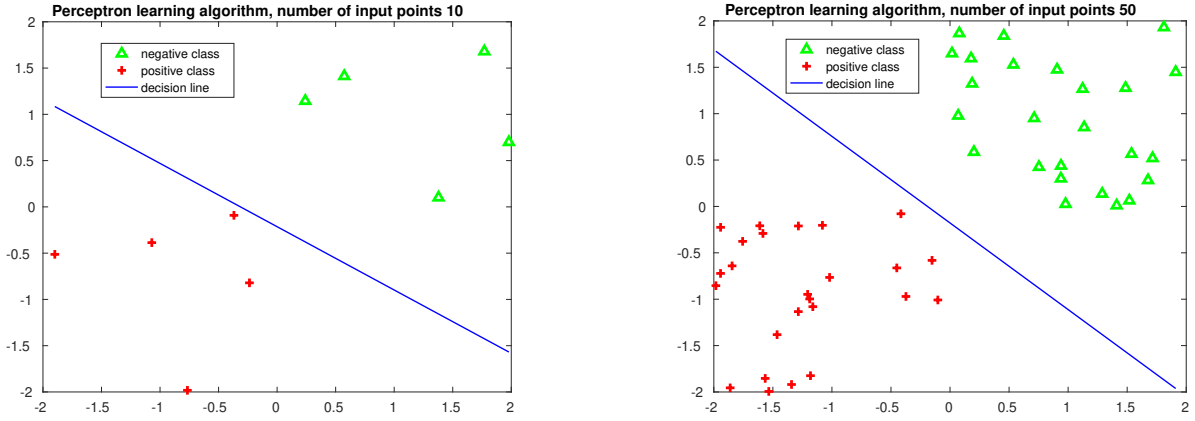
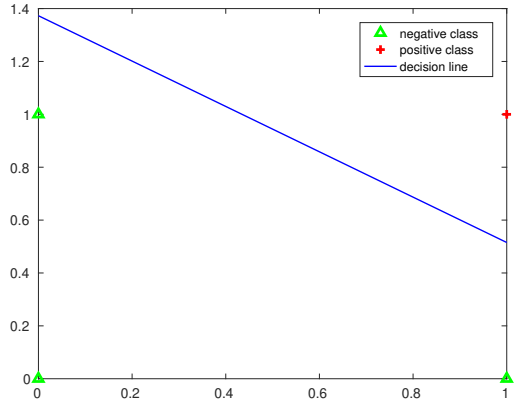
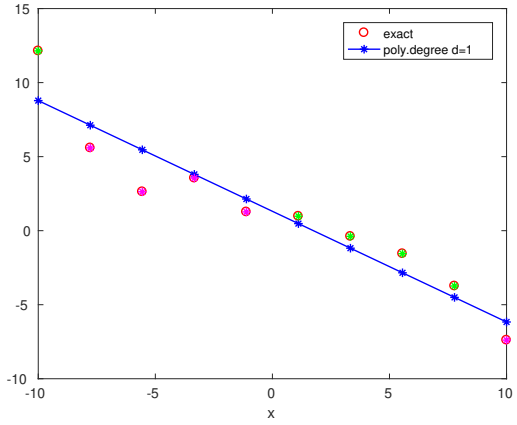


Figure: Examples of working perceptron learning algorithm which computes weights for decision line to separate two classes.

Implement in MATLAB all these classification algorithms and present decision lines for following training sets:



a)



b)

I) for randomly distributed data $y_i, i = 1, ..., m$ generated by

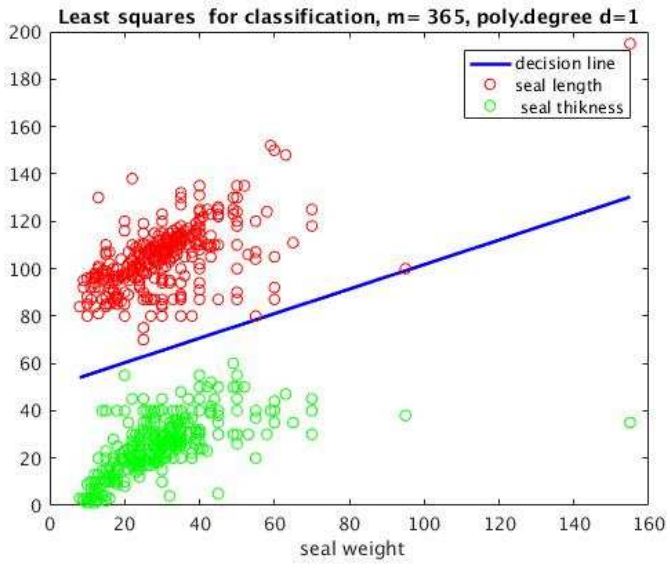
$$-1.2 + 0.5x + y = 0$$

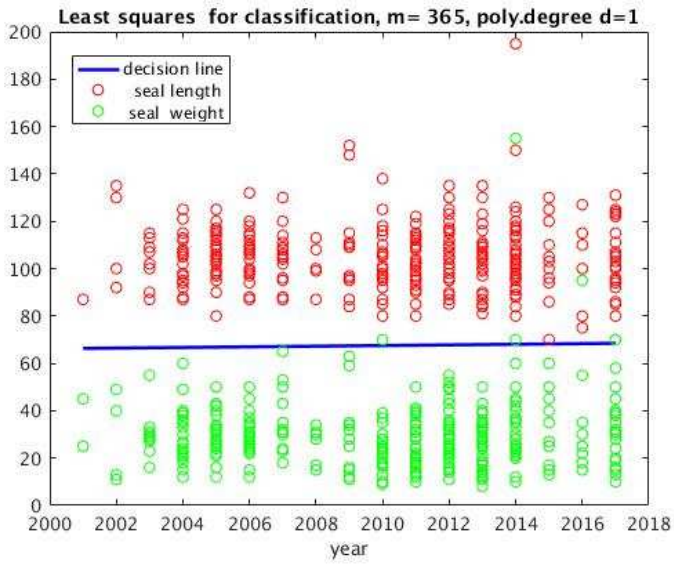
on the interval $x = [-10, 10]$. Generate random noise δ to data $y(x)$ using the formula

$$y_\delta(x) = y(x)(1 + \delta\alpha),$$

where $\alpha \in (-1, 1)$ is randomly distributed number and $\delta \in [0, 1]$ is the noise level. For example, if noise in data is 5%, then $\delta = 0.05$. Perform different experiments with different number of generated data $m > 0$ which you choose as you want. Using least squares approach you can get similar results as presented in Figure b).

II)





Generate your own data and try separate them. You can take experimental data from the link
https://archive.ics.uci.edu/ml/datasets.html
Or download *.xlsx file with data for grey seals and classify time-dependent data in this file for length and weight of seals. Another option is classification of weight-dependent data on length and thikness of seals.

Try answer to the following questions:

- Analize what happens with performance of perceptron learning algorithm if we take different learning rates $\eta \in (0, 1]$? For what values of η perceptron learning algorithm is more sensitive and when the iterative process is too slow?
- Analyze which one of the studied classification algorthms perform best and why?
- Try to explain why least squares approach can fail in the case when usual linear classifier is used.

Hints:

1. In this exercise we will assume that we will work in domains with two classes: positive class and negative class. We will assume that each training example $\mathbf{x}$ can have values 0 or 1 and we will label positive examples with $c(\mathbf{x}) = 1$ and negative with $c(\mathbf{x}) = 0$.
2. We will also assume that these classes are linearly separable. These two classes can be separated by a linear function of the form

$$\omega_0 + \omega_1 x + \omega_2 y = 0, \tag{35}$$

where x, y are Cartesian coordinates. Note that the equation (35) can be rewritten for the case of a linear least squares problem as

$$\omega_0 + \omega_1 x = -\omega_2 y \tag{36}$$

or as

$$-\frac{\omega_0}{\omega_2} - \frac{\omega_1}{\omega_2} x = y. \tag{37}$$

3. Suppose that we have measurements $y_i, i = 1, ..., m$ in (37) and our goal is to determine coefficients in (37) from these measurements. We can determine coefficients $\omega_0, \omega_1, \omega_2$ by solving the following least squares problem:

$$\min_{\omega} \|A\omega - y\|_2^2 \tag{38}$$

with $\omega = [\omega_1, \omega_2]^T = [-\frac{\omega_0}{\omega_2}, -\frac{\omega_1}{\omega_2}]^T$, rows in matrix A given by

$$[1, x_k], \quad k = 1, ..., m,$$

and vector $y = [y_1, ..., y_m]^T$.

4. The perceptron learning algorithm is the following:

Perceptron learning algorithm

Assume that two classes c(**x**)=1 and c(**x**)=0 are linearly separable.

Step 0. Initialize all weights ω_i in

$$\sum_{i=0}^{n} \omega_i x_i = 0$$

to small random numbers (note $x_0 = 1$). Choose an appropriate learning rate $\eta \in (0, 1]$.

- Step 1. For each training example $\mathbf{x} = (x_1, ..., x_n)$ whose class is $c(\mathbf{x})$ do:
  - (i) Assign $h(\mathbf{x}) = 1$ if

$$\sum_{i=0}^{n} \omega_i x_i > 0$$

    and assign $h(\mathbf{x}) = 0$ otherwise.
  - (ii) Update each weight using the formula

$$\omega_i = \omega_i + \eta \cdot [c(\mathbf{x}) - h(\mathbf{x})] \cdot x_i.$$

- Step 2. If $c(\mathbf{x}) = h(\mathbf{x})$ for **all training examples** stop, otherwise go to Step 1.

5. The decision line can be presented in Matlab for already computed weights by Perceptron learning algorithm using the formula (37).