The Delta Method

Random variable $Y = f(X)$, and we know $E(X)$ and $V(X)$.

What can we say about $E(Y)$ and $V(Y)$?

$\Rightarrow$ setup CI for Y based on CI for X

Linearize Y in X around $E(X)$

$$Y \approx f(E(X)) + f'(X)\Big|_{E(X)} (X - E(X)) + \underbrace{\text{higher order terms}}_{\text{ignore these}}$$

derivative evaluated @ $E(X)$

$$\Rightarrow E(Y) \approx f(E(X))$$

$$V(Y) \approx \left(f'(X)\Big|_{E(X)}\right)^2 V(X)$$

Example of application

Halflife $T = \dfrac{\ln 2}{\gamma}$ ← decay rate

$\hat{T} = \dfrac{\ln 2}{\hat{\gamma}}$ plug-in estimator

$$V(\hat{T}) \overset{\sim}{=} \left(-\frac{\ln 2}{\hat{\gamma}^2}\right)^2 V(\hat{\gamma}) \quad \Rightarrow \quad CI(T) = \left[\frac{\ln 2}{\hat{\gamma}} \pm q_N(1-\alpha/2)\left(\frac{\ln 2}{\hat{\gamma}^2}\right) SE(\hat{\gamma})\right]$$

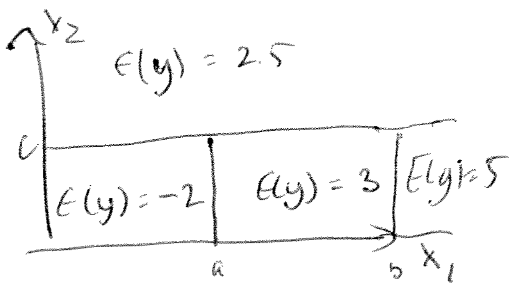
CART Classification & Regression Trees

Idea - partition X-space into rectangular regions where $E(y)$ takes on only one value

if regression tree; $E(Y \mid \text{region } k) = \mu_k$

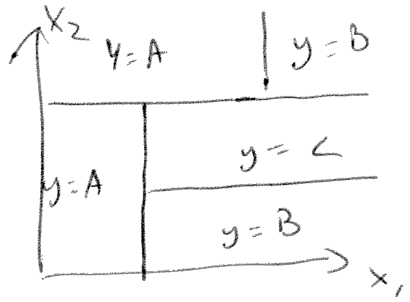
if classification tree; Y in region k = category l

Example



continuous / numerical y.

or



3-level y

The model looks like this

$$E(y \mid X \in R_k) = \mu_k$$

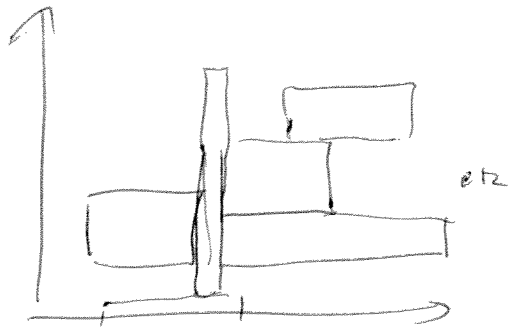
$$R_k = \{X_j > Th_j\} \times \{X_{j'} \le Th_{j'}\} \cdot \{X_{j''} \le Th_{j''}\}. \quad \text{etc}$$

Note can have same x-variable appear multiple times above

E.g. in example

$E(y) = 3$ region is $\{x_2 \le c\} \times \{x_1 > a\} \times \{x_1 \le b\}$

We can't build *any* type of rectangular regions . . .



. . . because the model space to search over

– which variables & which thresholds

is too large.

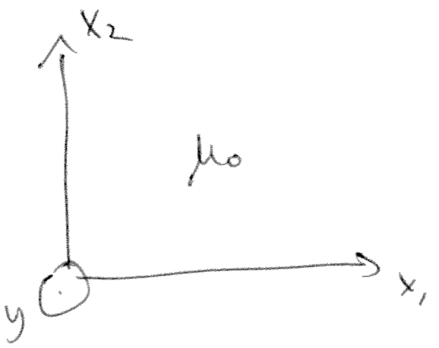
What CART does is use a *forward* search method.

The idea is to build the regions one variable ~~&~~ threshold at a time. Each step tries to make a newly created region as "pure" as possible

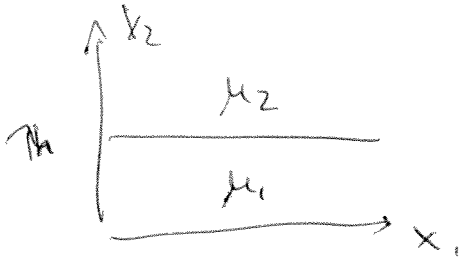
If y numerical → decrease overall MSE

If y categorical → decrease a "misclassification" error

Example



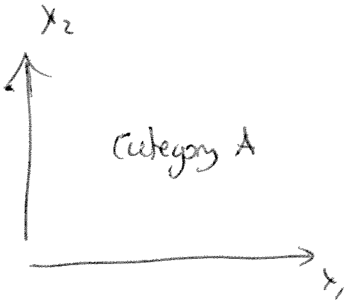
Before split $MSE_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_0)^2$



$$MSE_1 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mu_1 1\{x_{i2} \le \mathrm{Th}\} - \mu_2 1\{x_{i2} > \mathrm{Th}\}\right)^2$$

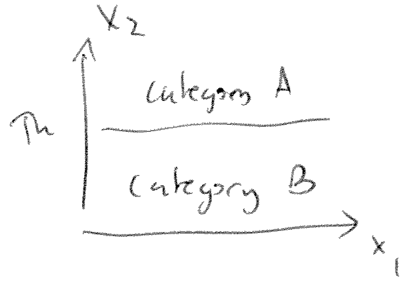
Find the x-variable and value for Th that minimizes the MSE

Example 2



Before split

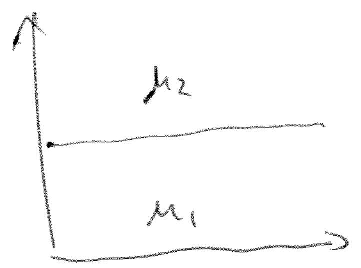
Misclass. error rate $= \frac{1}{n}\sum_{i=1}^{n} 1\{y_i \neq A\}$

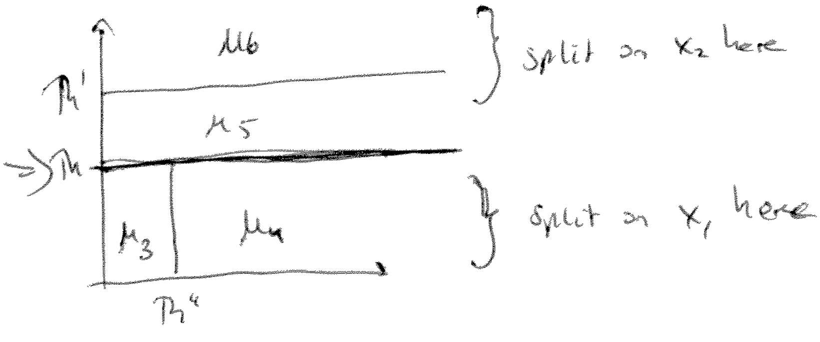


After

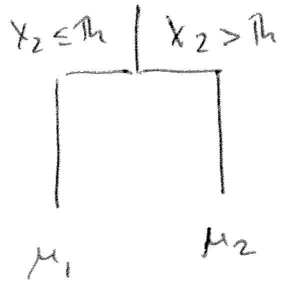
Misclass. error rate $= \frac{1}{n}\sum_{i=1}^{n}\Big(1\{x_{i2} > \mathrm{Th}\}\cdot 1\{y_i \neq A\} + 1\{x_{i2} \le \mathrm{Th}\}\cdot 1\{y_i \neq B\}\Big)$

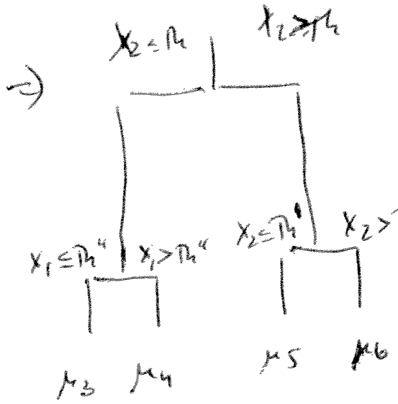
Next step — look at data in each region separately and find the ultimate split here





We can write these models in a tree format





The length of the branches are proportional to the reduction in MSE due to the split.

Here, the first split ($X_2 > Th$, $X_2 \le Th$) reduced the overall MSE much more than subsequent splits.

How many splits should we use?

- This is like choosing the number of variables in a regression model.

The most commonly used method is <u>Cross-validation</u>

Build a big tree, lots of splits. Then try "pruning" it by cutting branches. Evaluate the CV MSE or error rate after each pruning step and stop pruning when the CV error increases.

DEMO.

Cautionary remark.

CART is notoriously <u>unstable</u>

That is, if you run CART on a slightly different data set, the tree can look <u>very</u> <u>very</u> different – especially if x's are correlated.

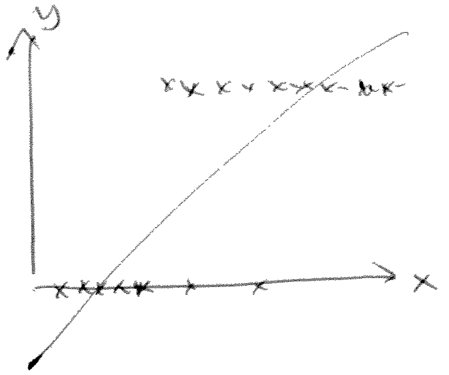
One of the best methods for prediction is to run CART on several bootstrap data sets (resample observations (x_i, y_i))

and use an <u>average</u> model, average prediction from all trees built.

This is called Bagging (but similar procedures are also used for non-tree models).

Generalized Linear Models (GLM)

What if $y \in \{0,1\}$, can we still run regression?



0-1 regression

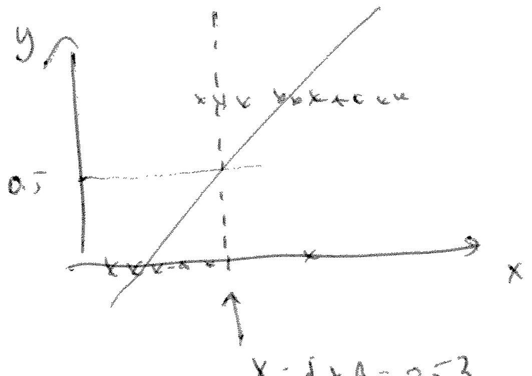
$E(y) = X\beta$

when $y \in \{0,1\}$ $E(y) = P(y=1)$

so interpretation of regression line is

$\boxed{P(y=1|X=x) = x\beta}$

Natural to use $p > 0.5$ as a cut-off for predicting $y=1$



$X : \{x\beta = 0.5\}$ is called the decision boundary.

This easily generalizes to more than one x, and we can run model selection etc.

Problem? $x\beta$ can be >1 and <0 so cannot really represent $p(y=1|X=x)$.

Logistic regression

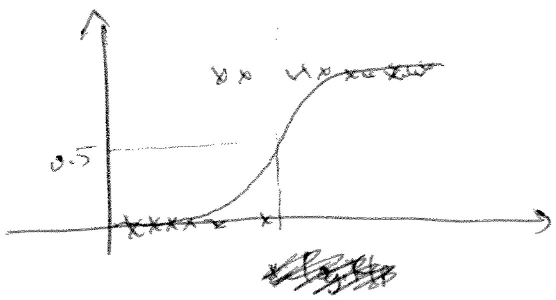
What if we don't assume $p(y=1|X=x)$ is linear in X but rather that some transformation of p is linear in x.

Example of a transformation that works, Logit

$$\text{logit}: \log \frac{p(y=1|X=x)}{1-p(y=1|X=x)} = \log \frac{p(y=1|X=x)}{p(y=0|X=x)} = X\beta$$

"log-odds"

$$\Rightarrow p(y=1|X=x) = \frac{e^{X\beta}}{1+e^{X\beta}} \in (0,1) \text{ always}$$



$$x: \frac{e^{X\beta}}{1+e^{X\beta}} = 0.5$$

General idea

$$E(y_i) = \mu_i \Rightarrow g(\mu_i) = x_i'\beta$$

link function!

Recap of models we now can work with

(1) LM $\quad y = X\beta + \varepsilon \quad \begin{cases} E(\varepsilon) = 0 \\ V(\varepsilon) = \sigma^2 I \\ \varepsilon \sim N(0, \sigma^2) \end{cases} \Rightarrow y \sim N(X\beta, \sigma^2 I)$

$\Rightarrow \begin{cases} \hat{\beta} = (X'X)^{-1}X'y \\ \hat{\sigma}^2 = \dfrac{RSS(\hat{\beta})}{n-p} \end{cases}$

(closed form solutions)

(2) NLM $\quad y = f(X;\beta) + \varepsilon \quad \begin{cases} E(\varepsilon) = 0 \\ V(\varepsilon) = \sigma^2 I \\ \varepsilon \sim N(0, \sigma^2 I) \end{cases} \Rightarrow y \sim N(f(X;\beta), \sigma^2 I)$

$\Rightarrow$ fit by iterative least squares (ILS)

$\rightarrow$ Careful w. starting values, convergence & confidence intervals

Can use F-test for model selection still.

(3) GLM $\quad y \sim f_Y(X;\beta)$ — allow for non-normal errors

distribution for Y not nec. normal

examples: Binomial, Poisson

Binomial : $y_i \sim Bin(m, \pi_i)$, $\quad \text{logit}(\pi_i) = x_i\beta$ (need $\pi_i \in (0,1)$)

Poisson : $y_i \sim Poi(\mu_i)$, $\quad \log(\mu_i) = x_i\beta$

(need $\mu_i > 0$)

Assumptions for GLM

- y_i independent observations (like uncorr. ε in LM)
- $E(y_i) = \mu_i$, $g(\mu_i) = \eta_i = x_i'\beta$

link function

η_i called the linear predictor.

- $V(y_i)$ related to μ_i in a known fashion

Example $y_i \sim \text{Poisson}(\mu_i) \Rightarrow \begin{cases} E(y_i) = \mu_i \\ V(y_i) = \mu_i \end{cases}$

$y_i \sim \text{Bin}(m, \pi_i) \Rightarrow \begin{cases} E(y_i) = m\pi_i \\ V(y_i) = m\pi_i(1-\pi_i) \end{cases}$

Also – assume $g()$ is the "correct" link (fits the data)

– and no outliers

Example, radioactive decay data

NLM $y_i = \alpha + \beta e^{-\gamma t_i} + \varepsilon_i$

Now $y_i \sim \text{Poi}(\mu_i)$, $\mu_i = \beta e^{-\gamma t_i} = e^{\log\beta - \gamma t_i} = e^{(\beta' + \gamma' t_i)} = e^{x_i'\beta}$

Linear predictor $\beta' + \gamma' t_i$, (β', γ') reparametrization of problem into what is called "canonical parameters"

Poisson is an example of an exponential family distribution (5)

— properties of GLM well understood here.

Definition

exp. family distribution $f(y \mid \theta, \varphi) = \exp\left[\underbrace{\frac{y\theta - b(\theta)}{a(\varphi)}}_{\text{like RSS}} + \underbrace{c(y, \varphi)}_{\substack{\text{normalizing constant that}\\ \text{does not depend on } \theta}} \right]$

θ: mean parameters ("β's")

φ: scale parameter ("σ")

→ can generalize to $f(y_i \mid \theta_i, \varphi)$

$\theta_i = \theta_i(x_i)$

<u>Example</u>. normal error

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}(y-\mu)^2\right\} \qquad \mu = X\beta$$

$$-\log f(y) = \frac{-1}{2\sigma^2}(y-\mu)^2 - \frac{1}{2}\log(2\pi\sigma^2)$$

$$= \frac{1}{\sigma^2}\left(y\mu - \frac{1}{2}\mu^2 - \frac{1}{2}y^2\right) - \frac{1}{2}\log 2\pi\sigma^2$$

$$\Rightarrow \begin{cases} \theta = \mu = X\beta \\ \varphi = \sigma^2 \end{cases}$$

that is $g(\mu) = \theta$ = identity link

$$\Rightarrow b(\theta) = \frac{1}{2}\mu^2$$

Poisson

$$f(y) = \frac{e^{-\mu}\mu^y}{y!}$$

$$\log f(y) = y\log\mu - \mu - \log y!$$

$$\Rightarrow \begin{cases} \theta = \log(\mu) & \Rightarrow \mu = e^\theta \quad \text{log-link} \\ (= X_i\beta) & \\ \varphi = 1 & \end{cases}$$

$$\Rightarrow b(\theta) = \mu = e^\theta$$

Binomial

$$f(y) = \binom{m}{y}\pi^y(1-\pi)^{1-y}$$

$$\log f(y) = \log\binom{m}{y} + y\log\pi + (1-y)\log(1-\pi)$$

$$= y\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi) + \log\binom{m}{y}$$

$$\Rightarrow \begin{cases} \theta = \log\left(\frac{\pi}{1-\pi}\right) \\ \varphi = 1 \end{cases}$$

WLS

Let's say $V(y_i) = \sigma_i^2$ not constant σ^2

$\Rightarrow V(y) = \sigma^2 V$ varying components. V non-diagonal — y's are correlated as well.

$\Rightarrow$ Use LS $\Rightarrow$ $\hat{\beta} = (X'X)^{-1}X'y$

$$E(\hat{\beta}) = \beta \;,\; V(\hat{\beta}) = \sigma^2 (X'X)^{-1}(X'VX)(X'X)^{-1}$$

WLS

$$\min_{\beta} \sum_{i=1}^{n} w_i (y_i - x_i'\beta)^2 = \sum_{i=1}^{n} (w_i^{1/2}(y_i - x_i'\beta))^2 = \| W^{1/2} y - W^{1/2} X\beta \|^2$$

$$= \|\tilde{y} - \tilde{X}\beta\|^2$$

transformed variables $\begin{cases} \tilde{y} = W^{1/2} y \\ \tilde{X} = W^{1/2} X \end{cases}$

$$\Rightarrow \hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} \qquad , \hat{y} = Hy = \tilde{X}'\hat{\beta}$$

$$= (X'WX)^{-1}X'Wy \qquad = \underbrace{\left(W^{1/2}X(X'WX)^{-1}X'W^{1/2}\right)}_{H} y$$

Now if $V(y) = \sigma^2 V$ and we use $W = V^{-1}$

$\Rightarrow V(\hat{\beta}) = \sigma^2 (X'WX)^{-1}$ best possible

That is, use weights $w_i \sim \dfrac{1}{V(y_i)}$