# MSA101/MVE187 2018 Lecture 10

Petter Mostad

Chalmers University

October 4, 2018

# Overview

- Some information theory.
- The EM algorithm.
- An example where the EM algorithm is used.

# The information of an event

We assume given a probability mass function $\pi(x)$ on a finite set.

- We want to define the "information" $h(x)$ in an event $x$. Requirements:
  - An event with probability 1 should have zero information.
  - The information should increase with decreasing probability $\pi(x)$.
  - The information in two independent events should be the sum of the information in each.
- We define $h(x) = -\log(\pi(x))$.
- When using the base 2 logarithm $\log_2$, information is measured in "bits". We however use the natural logarithm.

# Expected information: Entropy

- Define the entropy $H[X]$ of the random variable $X$ as the expected information:

$$H[X] = \sum_x h(x)\pi(x) = -\sum_x \pi(x)\log(\pi(x))$$

- Example: A uniform distribution on $n$ values has entropy $\log(n)$. This is the largest entropy possible for a distribution on $n$ values.
- Shannon's coding theorem: The entropy (using $\log_2$) is a lower bound on the expected number of bits needed to tranfer the information from $X$.

# (Differential) entropy for continuous distributions

- For any random variable $X$, its (differential) entropy is defined as

$$H[X] = \mathrm{E}\left[-\log(\pi(x))\right] = -\int_x \log(\pi(x))\pi(x)\,dx$$

- $H[X]$ may now be negative.
- Example: Assume $X \sim \mathrm{Normal}(\mu, \sigma^2)$. Then

$$\begin{aligned}\mathrm{E}\left[-\log(\pi(x))\right] &= \mathrm{E}\left[-\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{1}{2\sigma^2}(x-\mu)^2\right] \\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathrm{E}\left[(x-\mu)^2\right] = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}.\end{aligned}$$

- In fact, among all random variables $X$ with $\mathrm{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$, the normal has the largest entropy.

# Conditional entropy and mutual information

- The conditional entropy is defined as

$$H[Y|X] = \int \left[ \int \pi(y \mid x)(-\log(\pi(y \mid x)))\, dy \right] \pi(x)\, dx$$

- Show that

$$H[X, Y] = H[Y|X] + H[X].$$

- The mutual information is defined as

$$I[X, Y] = -\int \int \pi(x, y) \log \left( \frac{\pi(x)\pi(y)}{\pi(x, y)} \right) \, dx\, dy$$

- Show that

$$I[X, Y] = H[X] + H[Y] - H[X, Y]$$

# The Kullback-Leibler distance (relative entropy)

- For two densities $p(x)$ and $q(x)$ we define the Kullback-Leibler distance from $p$ to $q$ as

$$\mathrm{KL}[p||q] = -\int p(x) \log\left(\frac{q(x)}{p(x)}\right)\, dx$$

- Note that $\mathrm{KL}[p||q]$ is generally different from $\mathrm{KL}[q||p]$.
- However, it has the distance property that $\mathrm{KL}[p||q] \geq 0$ always, while $\mathrm{KL}[p||q] = 0$ if and only if $p = q$.
- The standard proof uses Jensen's inequality.
- Note that

$$\mathrm{KL}\left(\pi(x, y)||\pi(x)\pi(y)\right) = I[X, Y]$$

- Note that

$$\mathrm{KL}[p||q] = \mathrm{E}_p\left[-\log(q(x))\right] - H[X]$$

where $X$ is a random variable with density $p(x)$.

# Example

Assume $X \sim \mathrm{Normal}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{Normal}(\mu_Y, \sigma_Y^2)$.
Show by direct computation that

$$\mathrm{KL}\left[\pi_X || \pi_Y\right] = \frac{1}{2}\log(2\pi\sigma_Y^2) + \frac{\sigma_X^2}{2\sigma_Y^2} + \frac{1}{2\sigma_Y^2}(\mu_X - \mu_Y)^2 - \frac{1}{2}\log(2\pi\sigma_X^2) - \frac{1}{2}.$$

We see how the result is zero when the two distributions are identical.
We see how $\mathrm{KL}\left[\pi_X || \pi_Y\right] \neq \mathrm{KL}\left[\pi_Y || \pi_X\right]$ in general.

# The EM algorithm

- We want to find the $\theta$ maximizing the posterior $\pi(\theta \mid x)$; i.e., maximizing

$$\log\left(\pi(x \mid \theta)\pi(\theta)\right) = \log(\pi(x \mid \theta)) + \log(\pi(\theta))$$

- Assume we have a joint model $\pi(x, z \mid \theta)$ which includes augmented data $z$. We may then write, for any density $q(z)$,

$$\log(\pi(x \mid \theta)) + \log(\pi(\theta)) = \mathrm{KL}(q||\pi_z) + \mathcal{L}(q, \theta) + \log(\pi(\theta)) \qquad (1)$$

where

$$\mathcal{L}(q, \theta) = \int q(z) \log\left(\frac{\pi(x, z \mid \theta)}{q(z)}\right) dz$$

and

$$\mathrm{KL}(q||\pi_z) = -\int q(z) \log\left(\frac{\pi_z(z \mid x, \theta)}{q(z)}\right) dz$$

# The EM algorithm, cont.

- Fix $q(z) = \pi_z(z \mid x, \theta^{old})$ for some value $\theta^{old}$.
- With this $q(z)$, $\mathrm{KL}(q||\pi_z)$ will be zero when $\theta = \theta^{old}$ and positive for other $\theta$'s. THUS: If we find $\theta^{new}$ maximizing $\mathcal{L}(q,\theta) + \log(\pi(\theta))$, so that $\mathcal{L}(q,\theta^{new}) + \log(\pi(\theta^{new})) > \mathcal{L}(q,\theta^{old}) + \log(\pi(\theta^{old}))$, replacing $\theta^{old}$ with $\theta^{new}$ will increase the right side of Equation 1, and thus also the left side.
- Set $\theta^{old}$ to the value $\theta^{new}$ and start again from the first step above. Continue until convergence.
- Note that maximizing $\mathcal{L}(q,\theta) + \log(\pi(\theta))$ is the same as maximizing

$$\int q(z) \log\left(\pi(x, z \mid \theta)\right) dz + \log(\pi(\theta))$$

  where the left term is the expected full loglikelihood, taking the expectation over the density $q(z) = \pi_z(z \mid x, \theta^{old})$.
- E-step: Computing the expectation above. M-step: Maximizing.

# A simple example

We have data $x_1, \dots, x_n$, where we assume the following model, with a single parameter $\mu$: With probability 0.5, $x_i \sim \text{Normal}(0, 1)$ and with probability 0.5, $x_i \sim \text{Normal}(\mu, 1)$. We assume a flat prior on $\mu$.

- The likelihood can be written as

$$\pi(x_1, \dots, x_n \mid \mu) = \prod_{i=1}^{n} (0.5 \cdot \text{Normal}(x_i; 0, 1) + 0.5 \cdot \text{Normal}(x_i; \mu, 1))$$

- With the loglikelihood programmed numerically, we may
  - Optimize to find the maximum likelihood estimate $\hat{\mu}$ for $\mu$.
  - Simulate from the posterior, using, e.g., Metropolis Hastings.
- Instead, we may introduce *augmented* data $z_1, \dots, z_n$, where each $z_i$ has value 0 or 1, so that $z_i \sim \text{Bernoulli}(0.5)$ and $x_i \mid z_i \sim \text{Normal}(z_i \cdot \mu, 1)$. The full posterior may be written as

$$\pi(x_1, \dots, x_n, z_1, \dots, z_n, \mu) \propto \prod_{i=1}^{n} \pi(x_i \mid z_i, \mu) = \prod_{i=1}^{n} \text{Normal}(x_i; z_i \cdot \mu, 1)$$

- The augmented model may be used both for simulation (using Gibbs sampling) and for finding the maximum aposteriori value for $\mu$ using the EM-algorithm.

# A simple example: Using the EM algorithm

- First, find the complete data loglikelihood (or log posterior) which is (up to a constant)

$$l(\mu) = \sum_{i=1}^{n} -\frac{1}{2}(x_i - z_i \cdot \mu)^2$$

- Then, for a fixed value $\mu = \mu^{old}$, find the distribution $z_i \mid x_i, \mu^{old}$:

$$\begin{aligned} \pi(x_1, \ldots, x_n, \ldots, z_i = 0, \ldots, \mu^{old}) &= K \cdot \text{Normal}(x_i; 0, 1) \\ \pi(x_1, \ldots, x_n, \ldots, z_i = 1, \ldots, \mu^{old}) &= K \cdot \text{Normal}(x_i; \mu^{old}, 1) \end{aligned}$$

  Normalizing the distribution, we get

$$\begin{aligned} z_i \mid x_i, \mu^{old} &\sim \text{Bernoulli}(p_i)\,, \text{ where} \\ p_i &= \frac{\text{Normal}(x_i; \mu^{old}, 1)}{\text{Normal}(x_i; 0, 1) + \text{Normal}(x_i; \mu^{old}, 1)} \end{aligned}$$

- E step: Compute $\mathrm{E}_Z[l(\mu)]$. M step: Set $\mu^{new}$ as the parameter maximizing this function.

# A simple example continued

- The E step becomes

$$\begin{aligned} \mathrm{E}_Z[l(\mu)] &= \mathrm{E}_Z\left[\sum_{i=1}^{n} -\frac{1}{2}(x_i - z_i\mu)^2\right] \\ &= \mathrm{E}_Z\left[-\frac{1}{2}\sum_{i=1}^{n} x_i^2 - 2x_i z_i\mu + z_i^2\mu^2\right] \\ &= -\frac{1}{2}\sum_{i=1}^{n} x_i^2 - 2x_i\,\mathrm{E}_Z[z_i]\mu + \mathrm{E}_Z[z_i^2]\mu^2 \\ &= -\frac{1}{2}\sum_{i=1}^{n} x_i^2 - 2x_i p_i\mu + p_i\mu^2 \end{aligned}$$

- The M step becomes

$$\frac{\partial}{\partial\mu}\mathrm{E}_Z[l(\mu)] = -\frac{1}{2}\sum_{i=1}^{n}(-2x_i p_i + 2p_i\mu) = \sum_{i=1}^{n} x_i p_i - \mu\sum_{i=1}^{n} p_i = 0$$

resulting in $\mu^{new} = \left(\sum_{i=1}^{n} x_i p_i\right) / \left(\sum_{i=1}^{n} p_i\right)$.