MSA101/MVE187 2021 Lecture 12
Some Information Theory
The EM algorithm

Petter Mostad

Chalmers University

October 5, 2021

Overview

- Some information theory.
- The EM algorithm.
- A toy example.
- The Baum-Welsh algorithm as an example of EM.

The information of an event

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

- We want to define the "information" $h(U)$ in an event $U \subseteq S$. Requirements:
  - An event with probability 1 should have zero information.
  - The information should increase with decreasing probability $\pi(U)$.
  - The information in two independent events should be the sum of the information in each.
- We define $h(x) = -\log(\pi(x))$ for $x \in S$.
- 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:

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

- Note: $H[X]$ is always non-negative.
- 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 transfer 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 divergence (relative entropy)

- For two densities $p(x)$ and $q(x)$ we define the Kullback-Leibler divergence 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*.
- Jensen's inequality: If a function ψ is *convex*, then $\psi(\mathrm{E}[X]) \leq \mathrm{E}[\psi(X)]$.

The KL divergence

- Note that
$$\mathrm{KL}\left(\pi(x,y)||\pi(x)\pi(y)\right) = I[X,Y]$$
- Note that
$$\mathrm{KL}[p||q] = \mathrm{E}_p\left[-\log(q(x))\right] - H_p[X]$$
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.

Maximum posterior (MAP)

- The Maximal APosteriori (MAP): The value $\hat{\theta}$ that maximizes the posterior $\pi(\theta \mid \text{data})$.
- When the prior is flat, $\pi(\theta) \propto 1$, this corresponds to finding the maximum likelihood (ML) estimate for θ.
- Recall the advantages and disadvantages of using a single estimate instead of the full posterior.
- The MAP should be easy to compute when θ consists of all unknown variables: Just differentiate $\log(\pi(\theta \mid \text{data}))$, i.e. differentiate $\log(\pi(\text{data} \mid \theta)\pi(\theta))$.
- Much harder if the model also contains other unknown variables Z: Then $\pi(\theta \mid \text{data})$ is the marginal of $\pi(\theta, Z \mid \text{data})$ and much harder to maximize.
- The Expectation-Maximization (EM) algorithm comes to the rescue...

The EM algorithm

- We want to find the θ 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 $q(z)$,

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

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 θ^{old}.
- With this $q(z)$, $\mathrm{KL}(q||\pi_z)$ will be zero when $\theta = \theta^{old}$ and positive for other θ's. THUS: If we find θ^{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 θ^{old} with θ^{new} will increase the right side of Equation 1, and thus also the left side.
- Set θ^{old} to the value θ^{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 toy example

We have data $x_1, \ldots, x_n$, where we assume the following model, with a single parameter μ: 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 μ.

- The likelihood can be written as

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

- We now introduce *augmented* data $z_1, \ldots, z_n$, where each z_i has value 0 or 1, so that $z_i \sim \text{Bernoulli}(0.5)$ and $x_i \mid z_i \sim \text{Normal}(\mu z_i, 1)$. The full joint density may be written as

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

- One way to use this model is for finding the μ maximizing the posterior using the EM-algorithm.

A toy example: Using the EM algorithm

- First, find the complete data logposterior (which in our case is the same as the loglikelihood). It is (up to a constant)

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

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

A toy example continued

- The E step becomes

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

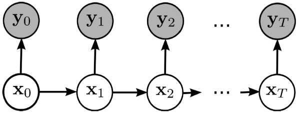
- The M step becomes

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

  resulting in $\mu^{new} = \left(\sum_{i=1}^{n} x_i p_i\right) / \left(\sum_{i=1}^{n} p_i\right)$.
- **Presentation break for R computations**

The Baum-Welch algorithm (as example of EM algorithm)

We consider an HMM where all the x_i have a finite state spaces



but where some of the parameters of the distributions $\pi(X_0)$, $\pi(X_i \mid X_{i-1})$, and $\pi(Y_i \mid X_i)$ are unknown. Objective: Given fixed values for the y_i, find maximum likelihood estimates for the parameters in the model.

- Note: If assuming flat priors the problem becomes that of computing the parameters maximizing the posterior, i.e., finding the MAP.
- Idea: Use the EM algorithm, with the values of the x_i as the augmented data.
- The E step of the EM algorithm is computed using (a small generalization of) the Forward-Backward algorithm.

The Baum-Welch algorithm: Example

For simplicity we assume each X_i can have values $1, \dots, M$. Let

$$\theta = (q, p) = ((q_1, \dots, q_M), (p_{11}, \dots, p_{MM}))$$

be the parameters we want to estimate, where

$$\begin{aligned} q_j &= \Pr(X_0 = j) \\ p_{jk} &= \Pr(X_i = k \mid X_{i-1} = j) \end{aligned}$$

The full loglikelihood given θ becomes

$$\begin{aligned} & \log\left(\pi(x_0, \dots, x_T, y_0, \dots, y_T \mid \theta)\right) \\ = \quad & \log\left(\pi(x_0 \mid \theta) \prod_{i=1}^{T} \pi(x_i \mid x_{i-1}, \theta) \prod_{i=0}^{T} \pi(y_i \mid x_i)\right) \\ = \quad & \log \pi(x_0 \mid \theta) + \sum_{i=1}^{T} \log \pi(x_i \mid x_{i-1}, \theta) + \sum_{i=0}^{T} \log \pi(y_i \mid x_i) \\ = \quad & C + \sum_{j=1}^{M} I(x_0 = j) \log q_j + \sum_{i=1}^{T} \sum_{j=1}^{M} \sum_{k=1}^{M} I(x_{i-1} = j) I(x_i = k) \log p_{jk} \end{aligned}$$

The Baum-Welch algorithm: Example continued

- In the E step, we would like to compute the expectation of the full loglikelihood under the distribution $\pi(x_0, \ldots, x_T \mid y_0, \ldots, y_T, \theta^{old})$ for some set of parameters θ^{old}.
- Thus we need to compute the expectations $\mathrm{E}\left[I(x_0 = j)\right]$ and $\mathrm{E}\left[I(x_{i-1} = j)I(x_i = k)\right]$ under this distribution.
- Fixing θ^{old}, we can use the Forward-Backward algorithm to compute the densities $\pi(x_i \mid y_0, \ldots, y_i)$ and $\pi(y_{i+1}, \ldots, y_T \mid x_i)$. Further we have that

$$\begin{aligned}
& \pi(x_i, x_{i+1} \mid y_0, \ldots, y_T) \\
\propto \quad & \pi(y_{i+1}, \ldots, y_T \mid x_i, x_{i+1})\pi(x_i, x_{i+1} \mid y_0, \ldots, y_i) \\
\propto \quad & \pi(y_{i+2}, \ldots, y_T \mid x_{i+1})\pi(y_{i+1} \mid x_{i+1})\pi(x_{i+1} \mid x_i)\pi(x_i \mid y_0, \ldots, y_i)
\end{aligned}$$

  making it possible to compute the joint posterior for x_i and x_{i+1} from these densities.

The Baum-Welch algorithm: Example continued

The algorithm can now be summed up as

- Choose starting parameters θ^{old}.
- Run the Forward-Backward algorithm on the Markov model with parameters θ^{old} to compute the numbers $\mathrm{E}\left[I(x_0 = j)\right]$ and $\mathrm{E}\left[I(x_{i-1} = j) I(x_i = k)\right]$.
- Find the θ maximizing the expected loglikelihood

  $$\sum_{j=1}^{M} \mathrm{E}\left[I(x_0 = j)\right] \log q_j + \sum_{i=1}^{T} \sum_{j=1}^{M} \sum_{k=1}^{M} \mathrm{E}\left[I(x_{i-1} = j) I(x_i = k)\right] \log p_{jk}$$

  In fact, we get

  $$\hat{q}_j = \mathrm{E}\left[I(x_0 = j)\right] \text{ and } \hat{p}_{jk} = \frac{\sum_{i=1}^{T} \mathrm{E}\left[I(x_{i-1} = j) I(x_i = k)\right]}{\sum_{k=1}^{M} \sum_{i=1}^{T} \mathrm{E}\left[I(x_{i-1} = j) I(x_i = k)\right]}$$

- Set $\theta^{old} = ((\hat{q}_1, \ldots, \hat{q}_M), (\hat{p}_{11}, \ldots, \hat{p}_{MM}))$ and iterate until convergence.
- **See implementation in R**