MDL exercises, eleventh handout (final obligatory homework exercises)

(due May 18th, 14:00)

1. **[1 point]** Let $f(x)$ be a density function on $[0, \infty)$ with fixed mean $1/\lambda$. Define $g(x) = \lambda e^{-\lambda x}$, the density function of the exponential distribution on the same domain and with the same mean. Show that $H(f)$ is maximized by choosing $f = g$, by evaluating $0 \leq D(f\|g)$.

   Solution:

$$\begin{aligned}
D(f\|g)\ln 2 &= \int_0^\infty (x) \ln\left(\frac{f(x)}{g(x)}\right) dx \\
&= -\int_0^\infty f(x)\ln g(x)dx + \int_0^\infty f(x)\ln f(x)dx \\
&= -E_f[\ln g(X)] - H(f) \\
&= -E_f[\ln\lambda - \lambda X] - H(f) \\
&= -\ln\lambda - \lambda E_f[X] - H(f) \\
&= -\ln\lambda - \lambda E_g[X] - H(f) \\
&= -E_g[\ln g(X)] - H(f) \\
&= H(g) - H(f),
\end{aligned}$$

   where we used that $E_f[X] = 1/\lambda = E_g[X]$. By nonnegativity of the Kullback-Leibler divergence, we see

$$D(f\|g)\ln 2 \geq 0 \Rightarrow H(g) - H(f) \geq 0,$$

   so indeed $H(f)$ is maximized by choosing $f = g$.

2. Jensen's inequality states that $E[f(X)] \geq f(E[X])$ for conve^x f. Use this inequality to find (a) **[1 point]** a lower bound on the entropy $H(P)$ for a distribution P on a finite sample space, and (b) **[1 point]** an upper bound on this entropy (Hint: for the upper bound, rewrite the entropy as $-\sum_x P(x)(f(1/P(x)))$ with $f \equiv -\log$). Compare this upper bound to the entropy for the uniform distribution on that sample space and for a nonuniform distribution on that space. In which case is the bound tighter?

   Solution:

   Let us denote $\mathcal{X}$ for the sample space. Then we have

$$\begin{aligned}
H(P) &= -\sum_{x\in\mathcal{X}} P(x)\log P(x) \\
&= E_P[-\log P(X)] \\
&\geq -\log\left(E[P(X)]\right) \\
&= -\log\left(\sum_{x\in\mathcal{X}} P(x)^2\right)
\end{aligned}$$

and

$$\begin{aligned}
H(P) &= -\sum_{x\in\mathcal{X}} P(x)\log P(x) \\
&= -\sum_{x\in\mathcal{X}} P(x)\cdot -\log\left(\frac{1}{P(x)}\right) \\
&= -E\left[-\log\left(\frac{1}{P(X)}\right)\right] \\
&\leq \log\left(E\left[\frac{1}{P(X)}\right]\right) \\
&= \log\left(\sum_{x\in\mathcal{X}} P(x)\frac{1}{P(x)}\right) \\
&= \log(|\mathcal{X}|).
\end{aligned}$$

Since the uniform distribution on $\mathcal{X}$ is a maximizer of the entropy, the bound is tighter for the uniform distribution. In fact, we have equality in this case:

$$\begin{aligned}
H(U) &= -\sum_{x\in\mathcal{X}} U(x)\log(U(x)) \\
&= \sum_{x\in\mathcal{X}} \frac{1}{|\mathcal{X}|}\log\left(|\mathcal{X}|\right) \\
&= \frac{1}{|\mathcal{X}|}\log\left(|\mathcal{X}|\right)\sum_{x\in\mathcal{X}} 1 \\
&= \log\left(|\mathcal{X}|\right).
\end{aligned}$$

3. [**1 point**] For two distributions P_0 and P_1 defined on the same space $\mathcal{X}$ with $P_0 \neq P_1$, let P_α be the α- mixture between P_0 and P_1, i.e. $P_\alpha(x) = (1-\alpha)P_0(x) + \alpha P_1(x)$. Show that the entropy $H(P_\alpha)$ is strictly concave as a function of $\alpha \in [0,1]$.

   Solution:

   The most straightforward way is to do this by twice differentiating to α. There is however a shorter way: fix an arbitrary distribution Q on $\mathcal{X}$ and consider the function

   $$f_Q(\alpha) := E_{P_\alpha}[-\log Q(X)] = (1-\alpha)E_{P_0}[-\log Q(X)] + \alpha E_{P_1}[-\log Q(X)].$$

   Obviously $f_Q(\alpha)$ is linear in α for each Q. Now fix $0 < \alpha' < 1$. Then $f_{P_{\alpha'}}(\alpha)$ is obviously a linear function of $\alpha \in [0,1]$ with $f_{P_{\alpha'}}(\alpha') = H(P_{\alpha'})$. Also, for all $\alpha \in [0,1]$, $H(P_\alpha) \leq f_{P_{\alpha'}}(\alpha)$, because $H(P) = \min_{\text{all } Q} E_P[-\log Q(X)]$. Hence, for all $0 < \alpha' < 1$ the entropy lies underneath its tangent at α';

but this means it must be a concave function (make a drawing). Where we use that $f_{P_{\alpha'}}(\alpha)$ is the tangent of $\alpha \mapsto H(P_\alpha)$ at α', because it is a linear function that touches the curve.

4. Consider the following three families of distributions. For each of these models, prove that they are an exponential family. HINT: you can show that a family is an exponential family by rewriting it in the exponential form $\frac{1}{Z(\beta)}e^{\beta\phi(x)}r(x)$ for some function $\phi(x)$.

   a) **[1 point]** The set of all distributions on $\{0,1\}$ with mean $E[X] = \theta$, for all $0 \le \theta \le 1$ (How is this set of distributions called?).

      Solution:

      This is the set of Bernoulli distributions. Let us denote this distribution by P_θ. We know that for $x \in \{0,1\}$, we can write it as

$$\begin{aligned} P_\theta(x) &= \theta^x(1-\theta)^{1-x} \\ &= (1-\theta)\left(\frac{\theta}{1-\theta}\right)^x \\ &= (1-\theta)e^{\ln\left(\frac{\theta}{1-\theta}\right)x}. \end{aligned}$$

      So we let $\beta = \ln\left(\frac{\theta}{1-\theta}\right)$, $Z(\beta) = \frac{1}{1-\theta} = 1 + e^\beta$, $\phi(x) = x$ and $r(x) = 1$. So this is indeed an exponential family.

   b) **[1 point]** The set of all normal distributions with a variance of one, for all means $\mu \in \mathbb{R}$.

      Solution:

      Let f be the density function of an arbitrary normal distribution with variance one and mean $\mu \in \mathbb{R}$. Then we have for $x \in \mathbb{R}$

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2}}.$$

      We group all terms that have dependency only on x together, similarly for μ, and all terms that have dependency on both:

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2}} \\ &= \frac{1}{\sqrt{2\pi}}e^{\frac{-x^2-\mu^2+2x\mu}{2}} \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}e^{\mu x}e^{-\frac{\mu^2}{2}}. \end{aligned}$$

      So we let $\beta = \mu$, $r(x) = e^{-\frac{x^2}{2}}$, $\phi(x) = x$ and $Z(\beta) = \sqrt{2\pi}e^{\mu^2/2} = \sqrt{2\pi}e^{\beta^2/2}$ and see that this is indeed an exponential family.

c) **[1 point]** The set of power law distributions, also known as the *Pareto family*: $P_\theta(n) = n^{-\theta} / \sum_{n=1}^{\infty} n^{-\theta}$ for $n \in \{1, 2, \ldots\}$ and $\theta > 1$.

Solution:

We see

$$n^{-\theta} = \frac{1}{n^\theta} = \frac{1}{(e^{\ln n})^\theta} = e^{-\theta \ln n}.$$

So we let $\beta = \theta, \phi(n) = -\ln n$, $r(n) = 1$ and $Z(\beta)$ simply as the normalizing term

$$\sum_{n=1}^{\infty} e^{\beta\phi(x)}.$$

So it indeed is an exponential family.

5. This question refers back to questions 4(a)-4(b).

a) **[1 point]** Is the distribution corresponding to θ in question 4(a) the maximum entropy distribution among *all* distributions on $\{0, 1\}$ with mean $E[X] = \theta$? Why (not)?

Solution:

It is the maximum entropy distribution, because it is the only distribution on $\{0, 1\}$ with mean θ. Let X be a random variable on $\{0, 1\}$ with mean θ. Then we see

$$E[X] = 0 \cdot \mathbb{P}(X = 0) + 1 \cdot \mathbb{P}(X = 1) = \mathbb{P}(X = 1).$$

Since any distribution on $\{0, 1\}$ is fully determined by the probability it gives to either 1 or 0, we see that the distribution is uniquely defined by its mean. There is, therefore, only one distribution on $\{0, 1\}$ with mean θ.

b) **[1 point]** Is the distribution corresponding to mean μ in question 4(b) the maximum entropy distribution among *all* distributions on $\mathbb{R}$ with mean μ? Why (not)?

Solution:

No it is not, the entropy increases with the variance, so for every μ we have $\mathcal{N}[\mu, 2]$ with higher entropy.