CPSC 486/586: Probabilistic Machine Learning	February 6, 2023
Lecture 7	
Lecturer: Andre Wibisono	*Scribe: Deyuan Li*

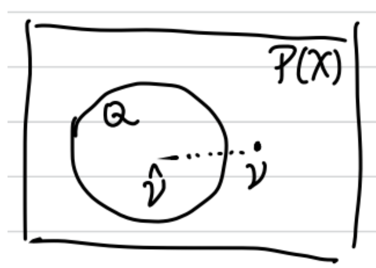
1 Approximating Probability Distributions (Cont'd)

We are continuing to ask how to approximate a distribution $\nu(x)$. In general, the distribution $\nu(x)$ can be very complicated, but we attempt to approximate ν with a similar distribution. Consider the space of probability distributions

$$\mathcal{P}(\mathcal{X}) = \left\{\rho : \mathcal{X} \to \mathbb{R} | \rho(x) \geq 0, \int_{\mathcal{X}} \rho(x)dx = 1\right\}.$$

Here $\mathcal{X}$ is the space of all values x can take on. If $\mathcal{X}$ is discrete such as $\mathcal{X} = \{1, \dots, k\}$, then $\Delta_k = \mathcal{P}(\mathcal{X}) = \{p = (p_1, \dots, p_k) \in \mathbb{R}^k : p_i \geq 0, \sum_{i=1}^k = 1\}$. In particular, this is a simplex in a $k-1$ dimensional subspace, and for $k = 2$, we have $\Delta_2 = (1-p, p) = \mathrm{Ber}(p)$.

Let $\mathcal{Q} \subset \mathcal{P}(\mathcal{X})$ be the space of nice (tractable) distributions. We want to approximate $\nu \in \mathcal{P}(\mathcal{X})$ by $\hat{\nu} \in \mathcal{Q}$.



For example, we can have

1. $\mathcal{Q} = \{\delta_x \in \mathcal{P}(\mathcal{X}) | x \in \mathcal{X}\}$. Then we can use the mode, $\hat{\nu} = \delta_{x^*}$, where $x^* = \arg\max_{x \in \mathcal{X}} \nu(x)$. This is the MAP estimator. We can also use the mean $\hat{\nu} = \delta_\mu$, where $\mu = \mathbb{E}_\nu[X]$.

2. We can approximate ν as Gaussian so we can let

$$\mathcal{Q} = \mathcal{G} = \{\rho = \mathcal{N}(m, C) : m \in \mathbb{R}^d, C \succ 0, C = C^\top \in \mathbb{R}^{d \times d}\}.$$

   We can use the Laplace Approximation (around the mode), which is $\boxed{\hat{\nu}_{\text{Lap}} = \mathcal{N}(x^*, (\nabla^2 f(x^*))^{-1})}$. In particular, if $\nu(x) \propto e^{-f(x)}$, then $\nabla^2 f = -\nabla^2 \log \nu$, and here, $x^* = \arg\min_x f(x)$.

2 Divergences

We now attempt to see if there are better approximations than the Laplace Distribution. Let

$$\mathcal{G} = \mathcal{N}(m, C) : m \in \mathbb{R}^d, C \in \mathbb{R}^{d\times d}, C \succ 0\}.$$

We can let $\hat{\nu} = \arg\min_{\rho\in\mathcal{G}} \mathcal{D}(\rho,\nu)$, where $\mathcal{D}(\rho,\nu)$ is any divergence satisfying $\mathcal{D}(\rho,\nu) \geq 0$ for all ρ,ν, and $\mathcal{D}(\rho,\nu) = 0$ if and only if $\rho = \nu$. The divergence is not necessary symmetric (i.e. $\mathcal{D}(\rho,\nu) \neq \mathcal{D}(\nu,\rho)$).

The following are some examples of divergences:

1. Total Variation Metric:

$$\mathrm{TV}(\rho,\nu) = \int_{\mathcal{X}} |\rho(x) - \nu(x)| dx = \int_{\mathcal{X}} \left|\frac{\rho(x)}{\nu(x)} - 1\right| \nu(x) dx = \mathbb{E}\left[\left|\frac{\rho}{\nu} - 1\right|\right] = \mathbb{E}_\nu \left[\left|\frac{\rho}{\nu} - \mathbb{E}_\nu\left[\frac{\rho}{\nu}\right]\right|\right].$$

   Note that $\mathbb{E}_\nu\left[\frac{\rho}{\nu}\right] = \int \frac{\rho(x)}{\nu(x)} \cdot \nu(x)dx = \int \rho(x)dx = 1$. This is actually a metric, meaning that the divergence is symmetric.

2. Chi-squared divergence:

$$\chi^2(\rho\|\nu) = \int_{\mathcal{X}} \left(\frac{\rho(x)}{\nu(x)} - 1\right)^2 \nu(x)dx = \mathrm{Var}_\nu\left(\frac{\rho}{\nu}\right) = \mathbb{E}_\nu\left[\left(\frac{\rho}{\nu} - 1\right)^2\right].$$

   This is not symmetric, and $\chi^2(\rho\|\nu) \neq \chi^2(\nu\|\rho)$.

3. KL-divergence:

$$\mathrm{KL}(\rho\|\nu) = \mathbb{E}_\rho\left[\log\frac{\rho}{\nu}\right] = \int \rho(x)\log\frac{\rho(x)}{\nu(x)}dx = \int \left(\frac{\rho(x)}{\nu(x)}\log\frac{\rho(x)}{\nu(x)}\right)\nu(x)dx = \mathbb{E}_\nu\left[\frac{\rho}{\nu}\log\frac{\rho}{\nu}\right].$$

   This is also not symmetric, and $\mathrm{KL}(\rho\|\nu) \neq \mathrm{KL}(\nu\|\rho)$.

In general, we can write $\mathcal{D}(\rho,\nu) = \mathbb{E}_\nu\left[\phi\left(\frac{\rho}{\nu}\right)\right]$ as some expectation. We see that for total variation, $\phi(x) = |x-1|$. For χ^2 divergence, $\phi(x) = (x-1)^2$. Finally for KL-Divergence, $\phi(x) = x\log x$.

It turns out that KL (Kullback-Leibler) divergence is generally the best. This is because it is equal to the relative entropy.

Lemma 1. *KL-divergence has the following properties:*

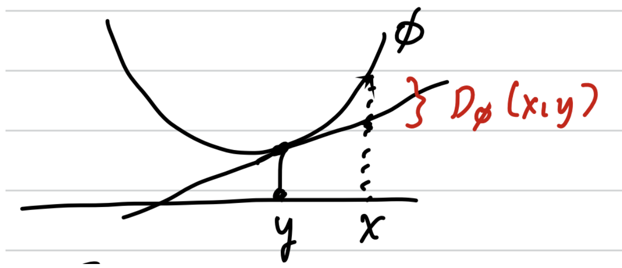
1. *$KL(\rho\|\nu) \geq 0$ for all ρ,ν*

2. *$KL(\rho\|\nu) = 0$ if and only if $\rho = \nu$.*

Proof. Recall by Jensen's Inequality that if ϕ is convex, then $\mathbb{E}[\phi(X)] \geq \phi(\mathbb{E}[X])$. Indeed, the **Bregman divergence** on a convex function ϕ is

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle.$$

This is the distance from $\phi(x)$ to the first-order approximation of ϕ centered at y, when evaluated at x. Since $\phi(x)$ is convex, $\mathcal{D}_\phi(x, y) \geq 0$.



Then using this fact about Bregman divergence,

$$\mathbb{E}[\mathcal{D}_\phi(X, \mathbb{E}[X])] \geq 0$$

since ϕ is convex. It turns out that

$$\mathbb{E}[\mathcal{D}_\phi(X, \mathbb{E}[X])] = \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X]),$$

meaning $\mathbb{E}[\phi(X)] \geq \phi(\mathbb{E}[X])$, which proves Jensen's Inequality. We now prove that KL-divergence is nonnegative. Recall that if we let $\phi : \mathbb{R}_+ \to \mathbb{R}$ be the function $\phi(R) = R \log R$, then

$$\mathrm{KL}(\rho \| \nu) = \mathbb{E}_\nu \left[\frac{\rho}{\nu} \log \frac{\rho}{\nu} \right] = \mathbb{E}_\nu \left[\phi \left(\frac{\rho}{\nu} \right) \right].$$

Note that $\phi(R)$ is convex, because $\phi'(R) = 1 + \log R$, so $\phi''(R) = \frac{1}{R} > 0$. By Jensen's Inequality, it follows that

$$\mathrm{KL}(\rho \| \nu) = \mathbb{E}_\nu \left[\phi \left(\frac{\rho}{\nu} \right) \right] \geq \phi \left(\mathbb{E}_\nu \left[\frac{\rho}{\nu} \right] \right) = \phi(1) = 1 \log 1 = 0.$$

Property 2 can also be checked. $\square$

KL-divergence is related to relative entropy $H(\rho)$.

Lemma 2. *KL-divegence is the Bregman divergence of the negative entropy.*

Proof. Without loss of generality, let $\mathcal{X} = \{1, \ldots, k\}$ for simplicity. The proof will similarly follow for all $\mathcal{X}$. Then for any $\rho = (p_1, \ldots, p_k) \in \mathcal{P}(\mathcal{X}) = \Delta_k$. Then the entropy is $H(\rho) = -\sum_{i=1}^k p_i \log p_i$. Note that $H(\rho)$ is a concave function of ρ, if we express $H : \Delta_k \to \mathbb{R}$. Indeed, in this case, we have

$$\nabla H(\rho) = \left(\frac{\partial H}{\partial p_i} \right)_{i=1}^k = (-1 - \log p_i)_{i=1}^k.$$

Then the Hessian is

$$\nabla^2 H(\rho) = \operatorname{diag}\left(-\frac{1}{p_i}\right) = \begin{pmatrix} -\frac{1}{p_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{1}{p_k} \end{pmatrix} \prec 0,$$

which implies that H is concave. Thus, $F(p) = -H(p) = \sum_{i=1}^k p_i \log p_i$ is convex. Then the Bregman divergence of the negative entropy is

$$\begin{aligned}
\mathcal{D}_F(p,q) &= F(p) - F(q) - \langle \nabla F(q), p - q \rangle \\
&= \sum_{i=1}^k p_i \log p_i - \sum_{i=1}^k q_i \log q_i - \sum_{i=1}^k (1 + \log q_i)(p_i - q_i) \\
&= \sum_{i=1}^k p_i \log \frac{p_i}{q_i} \\
&= \mathrm{KL}(p \| q),
\end{aligned}$$

which is the KL-divergence. Since Bregman divergence is always nonnegative, is also another proof that the KL-divergence is nonnegative. The proof follows similarly for continuous distributions where $\mathcal{X}$ is not necessarily discrete. In this case, we say $KL(p\|q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx$. □

As an example, we compute the entropy of a Gaussian distribution $\mathcal{N}(m, C)$. We find that

$$H(\rho) = -\mathbb{E}_\rho[\log \rho] = \mathbb{E}_\rho\left[\frac{1}{2}\|x - m\|_{C^{-1}}^2 + \frac{1}{2}\log\det(2\pi C)\right].$$

But in general, we know that the cyclic and linearity properties of the trace operator,

$$\begin{aligned}
\mathbb{E}_\rho[\|x - m\|_{C^{-1}}^2] &= \mathbb{E}[(x - m)^\top C^{-1}(x - m)] \\
&= \mathbb{E}[\mathrm{Tr}((x - m)^\top C^{-1}(x - m))] \\
&= \mathbb{E}[\mathrm{Tr}(C^{-1}(x - m)(x - m)^\top)] \\
&= \mathrm{Tr}(C^{-1}\mathbb{E}[(x - m)(x - m)^\top]) \\
&= \mathrm{Tr}(C^{-1} \cdot C) \\
&= \mathrm{Tr}(I_d) \\
&= d.
\end{aligned}$$

Thus, if $\rho = \mathcal{N}(m, C)$, then $H(\rho) = \frac{d}{2} + \frac{1}{2}\log\det(2\pi C) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det C$. Note that there is no dependence on m. Entropy is shift invariant and does not depend on mean. It measures the "shape" of the distribution and not the location.

For example, if $\rho = \mathcal{N}(0, \lambda I)$, then $H(\rho) = \frac{d}{2}\log\lambda + \frac{d}{2}\log(2\pi e)$. If $\lambda \ll 1$, then $H(p)$ is very negative. If $\lambda \gg 1$, then $H(p) \gg 1$.

Note that if the continuous case, the entropy is $H(\rho) = -\int_{\mathbb{R}^d} \rho \log \rho dx$ can be negative, but in the discrete case, the entropy $H(p) = -\sum_{i=1}^k p_i \log p_i \geq 0$ si always nonnegative.