Problem 1. KL and entropy

The Kullback-Leibler (KL) divergence of a distribution $p(\boldsymbol{x})$ from another distribution $q(\boldsymbol{x})$ is given by

$$D_{\mathrm{KL}}(p\|q) = -\int p(\boldsymbol{x}) \log\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}$$

Prove that $D_{\mathrm{KL}}(p\|q) \geq 0$.

Solution.

Theorem 1. Jensen's inequality *Suppose X is an integrable random variable, $f : \mathbb{R} \to \mathbb{R}$ is a concave function, such that $Y = f(X)$ is also integrable, then,*

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$$

Proof. Since f is concave, if $x, y \in \mathbb{R}$, we have

$$f(x) \leq f(y) + f'(y)(x - y)$$

Let $x = X$, $y = \mathbb{E}[X]$, then

$$f(X) \leq f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(X - \mathbb{E}[X])$$

This holds for all X. Thus, we could take the expectation on both side,

$$\mathbb{E}[f(X)] \leq \mathbb{E}[f(\mathbb{E}[X])] + \mathbb{E}[f'(\mathbb{E}[X])(X - \mathbb{E}[X])] = f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(\mathbb{E}[X] - \mathbb{E}[\mathbb{E}[X]]) = f(\mathbb{E}[X])$$

□

By Jensen's inequality, since log is concave, we have

$$\begin{aligned}
-D_{\mathrm{KL}}(p\|q) &= \int p(\boldsymbol{x}) \log\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} \\
&= \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\log\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right)\right] \\
&\leq \log\left(\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right]\right) \\
&= \log \int p(\boldsymbol{x}) \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \\
&= 0
\end{aligned}$$

You can also solve it by Log Sum Inequality or Gibbs' Inequality. ■

* If you have much time, also think about the following problems. Even if you don't think about them, we will cover them later in this course. These will not be covered in the recitation.

1. The Kullback-Leibler (KL) divergence of a distribution $p(\boldsymbol{x})$ from another distribution $q(\boldsymbol{x})$ is given by

$$D_{\mathrm{KL}}(p\|q) = -\int p(\boldsymbol{x}) \log\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right) d\boldsymbol{x}$$

Let $p(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{S})$. Calculate $D_{\mathrm{KL}}(p\|q)$.

Solution. The density functions for $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are

$$p(x) = \frac{1}{(2\pi)^{n/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \quad q(x) = \frac{1}{(2\pi)^{n/2}\det(\boldsymbol{S})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^\top \boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right)$$

Then,

$$
\begin{aligned}
D\left(p\|q\right) &= \int p(\boldsymbol{x})\log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)\mathrm{d}\boldsymbol{x} \\
&= \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\log p(\boldsymbol{x}) - \log q(\boldsymbol{x})\right] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\log\det\boldsymbol{\Sigma} - (\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \log\det\boldsymbol{S} + (\boldsymbol{x}-\boldsymbol{m})^\top\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right] \\
&= \frac{1}{2}\log\frac{\det\boldsymbol{S}}{\det\boldsymbol{\Sigma}} + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + (\boldsymbol{x}-\boldsymbol{m})^\top\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right] \\
&= \frac{1}{2}\log\frac{\det\boldsymbol{S}}{\det\boldsymbol{\Sigma}} + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\right) + \operatorname{tr}\left(\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{m})(\boldsymbol{x}-\boldsymbol{m})^\top\right)\right] \\
&= \frac{1}{2}\log\frac{\det\boldsymbol{S}}{\det\boldsymbol{\Sigma}} + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) + \operatorname{tr}\left(\boldsymbol{S}^{-1}\left(\boldsymbol{x}\boldsymbol{x}^\top - 2\boldsymbol{x}\boldsymbol{m}^\top + \boldsymbol{m}\boldsymbol{m}^\top\right)\right)\right] \\
&= \frac{1}{2}\log\frac{\det\boldsymbol{S}}{\det\boldsymbol{\Sigma}} - \frac{1}{2}n + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{S}^{-1}\left(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top - 2\boldsymbol{m}\boldsymbol{\mu}^\top + \boldsymbol{m}\boldsymbol{m}^\top\right)\right) \\
&= \frac{1}{2}\left(\log\frac{\det\boldsymbol{S}}{\det\boldsymbol{\Sigma}} - n + \operatorname{tr}\left(\boldsymbol{S}^{-1}\boldsymbol{\Sigma}\right) + \operatorname{tr}\left(\boldsymbol{\mu}^\top\boldsymbol{S}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^\top\boldsymbol{S}^{-1}\boldsymbol{m} + \boldsymbol{m}^\top\boldsymbol{S}^{-1}\boldsymbol{m}\right)\right) \\
&= \frac{1}{2}\left(\log\frac{\det\boldsymbol{S}}{\det\boldsymbol{\Sigma}} - n + \operatorname{tr}\left(\boldsymbol{S}^{-1}\boldsymbol{\Sigma}\right) + (\boldsymbol{m}-\boldsymbol{\mu})^\top\boldsymbol{S}^{-1}(\boldsymbol{m}-\boldsymbol{\mu})\right)
\end{aligned}
$$

■

2. The entropy of a distribution $p(\boldsymbol{x})$ is given by

$$H(p) = -\int p(\boldsymbol{x})\log p(\boldsymbol{x})d\boldsymbol{x}.$$

Calculate $H(p)$ where $p(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Solution. The density function for $p(\boldsymbol{x})$ is

$$p(x) = \frac{1}{(2\pi)^{n/2}\det(\boldsymbol{\Sigma})^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Then,

$$
\begin{aligned}
H(x) &= -\int p(\boldsymbol{x})\log p(\boldsymbol{x})d\boldsymbol{x} \\
&= -\mathbb{E}\left[\log\left(\frac{1}{(2\pi)^{n/2}\det(\boldsymbol{\Sigma})^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right)\right] \\
&= -\mathbb{E}\left[-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Sigma}) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \\
&= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}\mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \\
&= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}\mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\right)\right] \\
&= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\right]\right) \\
&= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) \\
&= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{I}\right) \\
&= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}n
\end{aligned}
$$

■

Problem 2. Ridge regression ([HaTF] Ex. 3.29)

Recall that in a ridge regression we minimize $\frac{1}{2}\|\boldsymbol{r} - \boldsymbol{X}\boldsymbol{w}\|^2 + \frac{1}{2}\lambda\|\boldsymbol{w}\|^2$. Suppose we run a ridge regression with parameter λ on a single variable x and get coefficient w (so the data matrix $\boldsymbol{X}$ is $N \times 1$, which can be denoted as a vector $\boldsymbol{x} \in \mathbb{R}^N$). We now include an exact copy $x^* = x$ and refit our ridge regression. Show that both coefficients are identical, and derive their value. Show in general that if m copies of a variable x_j are included in a ridge regression, their coefficients are all the same.

Solution. Let $\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{r} - \boldsymbol{X}\boldsymbol{w}\|^2 + \frac{1}{2}\lambda\|\boldsymbol{w}\|^2$, to find its minimum, we take the derivative,

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = -\boldsymbol{X}^\top(\boldsymbol{r} - \boldsymbol{X}\boldsymbol{w}) + \lambda\boldsymbol{w}$$

By setting the derivative to 0, we have

$$\boldsymbol{w} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}$$

For $\boldsymbol{X} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is a scaler. We have

$$\hat{w} = \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}}{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda}$$

If we include a copy of $\boldsymbol{X}$, the target function turned to

$$\underset{w_1, w_2}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{r} - \boldsymbol{X}w_1 - \boldsymbol{X}w_2\|^2 + \frac{1}{2}\lambda\|w_1\|^2 + \frac{1}{2}\lambda\|w_2\|^2$$

Let

$$\mathcal{L}(w_1, w_2) = \frac{1}{2}\|\boldsymbol{r} - \boldsymbol{X}w_1 - \boldsymbol{X}w_2\|^2 + \frac{1}{2}\lambda\|w_1\|^2 + \frac{1}{2}\lambda\|w_2\|^2$$

Take derivative with respect to w_1 and w_2

$$\begin{aligned}\frac{\partial \mathcal{L}(w_1, w_2)}{\partial w_1} &= -\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{r} - \boldsymbol{X}w_1 - \boldsymbol{X}w_2) + \lambda w_1 \\ \frac{\partial \mathcal{L}(w_1, w_2)}{\partial w_2} &= -\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{r} - \boldsymbol{X}w_1 - \boldsymbol{X}w_2) + \lambda w_2\end{aligned}$$

Set them to 0 and solve the system of equations, we have

$$\begin{bmatrix}\hat{w}_1 \\ \hat{w}_2\end{bmatrix} = \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}}{2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda}\begin{bmatrix}1 \\ 1\end{bmatrix}$$

They are the same.

If we have m copies of $\boldsymbol{X}$, the target function become

$$\underset{\boldsymbol{w}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{r} - \boldsymbol{X}\sum_{j=1}^m w_j\right\|^2 + \frac{1}{2}\lambda\sum_{j=1}^m\|w_j\|^2$$

Let

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\left\|\boldsymbol{r} - \boldsymbol{X}\sum_{j=1}^m w_j\right\|^2 + \frac{1}{2}\lambda\sum_{j=1}^m\|w_j\|^2$$

Take derivative with respect to w_k for $1 \le k \le m$

$$\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial w_k} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\sum_{j=1}^m w_j - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{r} + \lambda w_k$$

Set it to 0 and we have

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\sum_{j=1}^m \hat{w}_j + \lambda\hat{w}_k = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}$$

We could see that if we change k to another k', the solution will be the same. Thus, all $\hat{w}_k$ are the same

$$\hat{w}_k = \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}}{m\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda}, \quad 1 \le k \le m$$

■

Problem 3. Elastic net ([HaTF] Ex. 3.30)

Consider the elastic-net optimization problem:

$$\min_{\boldsymbol{w}} \|\boldsymbol{r}-\boldsymbol{X}\boldsymbol{w}\|^2+\lambda\left[\alpha\|\boldsymbol{w}\|^2+(1-\alpha)\|\boldsymbol{w}\|_1\right]$$

Show how one can turn this into a lasso problem using an augmented version of $\boldsymbol{X}$ and r :

$$\tilde{\boldsymbol{X}}=\begin{bmatrix}\boldsymbol{X}\\ \gamma\boldsymbol{I}\end{bmatrix}\ \text{, and}\quad \tilde{\boldsymbol{r}}=\begin{bmatrix}\boldsymbol{r}\\ \boldsymbol{0}\end{bmatrix}$$

Solution. Suppose $\boldsymbol{r}\in\mathbb{R}^N$, $\boldsymbol{X}\in\mathbb{R}^{N\times(D+1)}$, and $\boldsymbol{w}\in\mathbb{R}^{(D+1)}$. Then,

$$\tilde{\boldsymbol{X}}=\begin{bmatrix}\boldsymbol{X}\\ \gamma\boldsymbol{I}_{D+1}\end{bmatrix}\in\mathbb{R}^{(N+D+1)\times(D+1)},\quad \text{and } \tilde{\boldsymbol{r}}=\begin{bmatrix}\boldsymbol{r}\\ \boldsymbol{0}_{D+1}\end{bmatrix}\in\mathbb{R}^{(N+D+1)}$$

Thus, the lasso problem of $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{r}}$ is,

$$\left\|\tilde{\boldsymbol{r}}-\tilde{\boldsymbol{X}}\boldsymbol{w}\right\|^2+\tilde{\lambda}\|\boldsymbol{w}\|_1=\left\|\begin{matrix}\boldsymbol{r}-\boldsymbol{X}\boldsymbol{w}\\ \gamma\boldsymbol{w}\end{matrix}\right\|^2+\tilde{\lambda}\|\boldsymbol{w}\|_1=\|\boldsymbol{r}-\boldsymbol{X}\boldsymbol{w}\|^2+\gamma^2\|\boldsymbol{w}\|^2+\tilde{\lambda}\|\boldsymbol{w}\|_1$$

To make it a lasso problem, we need $\gamma^2=\alpha\lambda$, $\gamma=\sqrt{\lambda\alpha}$, and $\tilde{\lambda}=\lambda(1-\alpha)$.

Thus, to solve the elastic-net optimization problem, we first augment $\boldsymbol{X}$, $\boldsymbol{r}$ with $\gamma=\sqrt{\lambda\alpha}$, and solve the lasso problem of $\tilde{\boldsymbol{X}}$, $\tilde{\boldsymbol{r}}$ with $\tilde{\lambda}=\lambda(1-\alpha)$. ■

Problem 4. Kernel

1. Suppose $K(\boldsymbol{x})\geq 0$ and $\int_{\mathbb{R}^d}K(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=1$. Show that the kernel estimator

$$\hat{p}(\boldsymbol{x})=\frac{1}{Nh^d}\sum_{n=1}^{N}K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right)$$

is a density.

Solution. Suppose we have $\mathcal{X}=\{\boldsymbol{x}_n\}_{n=1}^N$, and we define,

$$\hat{f}(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}=\frac{1}{N}\mathbb{1}_{\mathcal{X}}(\boldsymbol{x})=\begin{cases}\frac{1}{N} & \text{if } \boldsymbol{x}\in\mathcal{X}\\ 0 & \text{if } \boldsymbol{x}\notin\mathcal{X}\end{cases},\quad \hat{K}(\boldsymbol{u})=\frac{1}{h^d}K(\frac{\boldsymbol{u}}{h})$$

Then, the convolution of $\hat{f}$ and $\hat{K}$ is,

$$\begin{aligned}\left(\hat{f}*\hat{K}\right)(\boldsymbol{x})&=\int_{\mathbb{R}^d}\hat{f}(\boldsymbol{u})\hat{K}(\boldsymbol{x}-\boldsymbol{u})\ \mathrm{d}\boldsymbol{u}\\ &=\frac{1}{N}\sum_{n=1}^{N}\hat{K}(\boldsymbol{x}-\boldsymbol{x}_n)\\ &=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^d}K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right)\\ &=\frac{1}{Nh^d}\sum_{n=1}^{N}K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right)=\hat{p}(\boldsymbol{x})\end{aligned}$$

Then,

$$\begin{aligned}\int_{\mathbb{R}^d} \hat{p}(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} &= \int_{\mathbb{R}^d} \left(\hat{f} * \hat{K}\right)(\boldsymbol{x})\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \hat{f}(\boldsymbol{u})\hat{K}(\boldsymbol{x}-\boldsymbol{u})\ \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{x}\\ &= \int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} \hat{K}(\boldsymbol{x}-\boldsymbol{u})\ \mathrm{d}\boldsymbol{x}\right)\hat{f}(\boldsymbol{u})\ \mathrm{d}\boldsymbol{u}\\ &= \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{u})\ \mathrm{d}\boldsymbol{u}\\ &= 1\end{aligned}$$

■

2. Suppose $K(\boldsymbol{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left[-\frac{\|\boldsymbol{x}\|^2}{2}\right]$. Show that each $K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right)$ can be written as a product of d univariate kernels.

Solution. Suppose $\boldsymbol{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(d)} \end{bmatrix}$, and $\boldsymbol{x}_n = \begin{bmatrix} x_n^{(1)} \\ x_n^{(2)} \\ \vdots \\ x_n^{(d)} \end{bmatrix}$.

The univariate Gaussian kernel is

$$K_0(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

Then,

$$\begin{aligned} K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right) &= \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left[-\frac{\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right\|^2}{2}\right]\\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left[-\frac{\|\boldsymbol{x}-\boldsymbol{x}_n\|^2}{2h^2}\right]\\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left[-\frac{\left(x^{(1)}-x_n^{(1)}\right)^2 + \left(x^{(2)}-x_n^{(2)}\right)^2 + \ldots + \left(x^{(d)}-x_n^{(d)}\right)^2}{2h^2}\right]\\ &= \prod_{k=1}^d \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(x^{(k)}-x_n^{(k)}\right)^2}{2h^2}\right)\\ &= \prod_{k=1}^d K_0\left(\frac{x^{(k)}-x_n^{(k)}}{h}\right)\end{aligned}$$

■

Problem 5. Smoother ([HaTF] Ex.6.8)

Suppose for continuous response Y and predictor X we model the joint density of X, Y using a multivariate Gaussian kernel estimator. This means that

$$\hat{p}(x,y) = \frac{1}{Nh^2}\sum_{n=1}^N K_h(x-x_n)\, K_h(y-y_n)$$

where $K_h(x) = K(x/h)$ and K is the Gaussian kernel. (cf. Problem 4 above.) Show that the conditional mean $\mathbb{E}[Y \mid X]$ derived from this estimate is a Nadaraya-Watson estimator.

Solution. By the definition of the conditional expectation,

$$\mathbb{E}[Y \mid X] = \int p(y \mid x) y\ \mathrm{d}y = \frac{\int p(x,y)y\ \mathrm{d}y}{p(x)}$$

The marginal distribution is,

$$\hat{p}(x) = \int \hat{p}(x, y) \ \mathrm{d}y = \frac{1}{Nh} \sum_{n=1}^{N} K_h\left(x - x_n\right)$$

Therefore,

$$\begin{aligned}
\mathbb{E}[Y \mid X] &= \int p(y \mid x) y \ \mathrm{d}y \\
&= \frac{\int p(x, y) y \ \mathrm{d}y}{p(x)} \\
&= \frac{\int \frac{1}{Nh^2} \sum_{n=1}^{N} K_h\left(x - x_n\right) K_h\left(y - y_n\right) y \ \mathrm{d}y}{\frac{1}{Nh} \sum_{n=1}^{N} K_h\left(x - x_n\right)} \\
&= \frac{\sum_{n=1}^{N} K_h\left(x - x_n\right)}{h \sum_{n=1}^{N} K_h\left(x - x_n\right)} \left(\int K_h\left(y - y_n\right) y \ \mathrm{d}y \right) \\
&= \frac{\sum_{n=1}^{N} K_h\left(x - x_n\right)}{h \sum_{n=1}^{N} K_h\left(x - x_n\right)} \left(\int K_h\left(y - y_n\right)\left(y - y_n\right) \ \mathrm{d}y + \int K_h\left(y - y_n\right) y_n \ \mathrm{d}y \right) \\
&= \frac{\sum_{n=1}^{N} K_h\left(x - x_n\right)}{h \sum_{n=1}^{N} K_h\left(x - x_n\right)} \left(0 + h y_n\right) \\
&= \frac{\sum_{n=1}^{N} K_h\left(x - x_n\right) y_i}{\sum_{n=1}^{N} K_h\left(x - x_n\right)}
\end{aligned}$$

Thus, it is a Nadaraya-Watson estimator. ■

Problem 6. EM (Midterm exam, Fall’21, Problem #-1)

Hilbert owns a PS5, an Xbox and a Switch (three different gaming systems). On each system there is a game. The outcome of each game is either win (“W”) or loss (“L”). Every day, he plays the Switch game. If he wins, he continues to play the PS5 game and records the outcome of the PS5 game; otherwise, he continues to play the Xbox game and records the outcome of the Xbox game. The outcomes he recorded for the last ten days of March are as follows:

W W L W W L W W L L

Suppose the event on each day is independent. Denote the (unknown) probabilities of ”W” for the PS5, the Xbox and the Switch games by p, q, π, respectively. Let y denote the random variable representing the final outcome, so that $y = 1$ if “W” is recorded, and $y = 0$ if “L” is recorded.

1. Suppose we use an expectation-maximization (EM) algorithm to find the maximum-likelihood solution of p, q, π. In plain language, describe what is the latent variable and the values it can take.

   Solution. The latent variable is the result of the Switch Game. It takes value in $\{0, 1\}$. ■

2. For the final outcome y, consider its parametric likelihood $p(y \mid p, q, \pi)$. Is it true or false that $p(y \mid p, q, \pi) = \sum_z (p(z \mid p, q, \pi) + p(y \mid z, p, q, \pi))$, where the summation is over all possible values of the latent variable?

   Solution. False. It is multiply: $p(y \mid p, q, \pi) = \sum_z p(z \mid p, q, \pi) p(y \mid z, p, q, \pi)$ ■

3. Write $p(y \mid p, q, \pi)$ as a function of y, p, q, π.

   Solution. Since π is the probability of wining the Switch game, we have

$$p(z = 1 \mid p, q, \pi) = \pi, \quad p(z = 0 \mid p, q, \pi) = 1 - \pi$$

If he wins the Switch game, he will play the PS5 with p wining rate, thus

$$p(y \mid z=1, p, q, \pi) = p^y (1-p)^{1-y}$$

If he loses the Switch game, he will play the Xbox with q wining rate, thus

$$p(y \mid z=0, p, q, \pi) = q^y (1-q)^{1-y}$$

Therefore, we have

$$\begin{aligned} p(y \mid p, q, \pi) &= \sum_{z \in \{0,1\}} p(z \mid p, q, \pi) p(y \mid z, p, q, \pi) \\ &= p(z=0 \mid p, q, \pi) p(y \mid z=0, p, q, \pi) + p(z=1 \mid p, q, \pi) p(y \mid z=1, p, q, \pi) \\ &= (1-\pi) q^y (1-q)^{1-y} + \pi p^y (1-p)^{1-y} \end{aligned}$$

■

4. Using the data recorded on the last ten days of March, and the initial values

$$\left(p^{(0)}, q^{(0)}, \pi^{(0)}\right) = (0.5, 0.5, 0.5)$$

implement the EM algorithm for one E-step and one M-step. Calculate $\left(p^{(1)}, q^{(1)}, \pi^{(1)}\right)$.

Solution.

To help with understanding, we write this EM as the form of GMM model in the lecture. This is a mixture of univariate Bernoulli. Let $\tilde{\boldsymbol{z}}$ be a one hot vector, and $\tilde{z}_k = 1$ implies the choice of the k-th cluster ($K = 2$ in our case). The marginal distribution over $\tilde{\boldsymbol{z}}$ is given by,

$$p(\tilde{z}_k = 1) = \tilde{\pi}_k, \text{ where } 0 \le \tilde{\pi}_k \le 1, \sum_{k=1}^{K} \tilde{\pi}_k = 1$$

Then, we can write

$$p(\tilde{\boldsymbol{z}}) = \prod_{k=1}^{K} \tilde{\pi}_k^{\tilde{z}_k}, \quad p(y \mid \tilde{z}_k = 1) = \mathcal{B}(y \mid \mu_k), \quad p(y \mid \tilde{\boldsymbol{z}}) = \prod_{k=1}^{K} \mathcal{B}(y \mid \mu_k)^{\tilde{z}_k}$$

where $\mathcal{B}(y \mid \mu_k) = \mu_k^y (1-\mu_k)^{(1-y)}$ is the univariate Bernoulli.

Therefore, the likelihood is

$$p(y) = \sum_{\tilde{\boldsymbol{z}}} p(\tilde{\boldsymbol{z}}) p(y \mid \tilde{\boldsymbol{z}}) = \sum_{k=1}^{K} \tilde{\pi}_k \mathcal{B}(y \mid \mu_k)$$

In order to derive the EM algorithm, we first write down the complete-data log likelihood function,

$$\log p(\boldsymbol{y}, \tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{\pi}}, \boldsymbol{\mu}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \tilde{z}_{nk} \left\{ \log \tilde{\pi}_k + y_n \log \mu_k + (1-y_n) \log(1-\mu_k) \right\}$$

Then, we take the expectation of the complete-data log likelihood with respect to the posterior distribution of the latent variables

$$\mathbb{E}_{\boldsymbol{z}}[\log p(\boldsymbol{y}, \tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{\pi}}, \boldsymbol{\mu})] = \underbrace{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\tilde{z}_{nk})\left\{\log\tilde{\pi}_k + y_n\log\mu_k + (1-y_n)\log(1-\mu_k)\right\}}_{\mathcal{L}}$$

In the E-step, we evaluate the responsibility term by Bayes' theorem,

$$\gamma(\tilde{z}_{nk}) = \mathbb{E}[\tilde{z}_{nk}] = \frac{\sum_{\tilde{z}_{nk}}\tilde{z}_{nk}[\tilde{\pi}_k p(y_n \mid \mu_K)]^{\tilde{z}_{nk}}}{\sum_{\tilde{z}_{nj}}[\tilde{\pi}_j p(y_n \mid \mu_K)]^{\tilde{z}_{nj}}} = \frac{\tilde{\pi}_k p(y_n \mid \mu_k)}{\sum_{j=1}^{K}\tilde{\pi}_j p(y_n \mid \mu_j)}$$

In the M-step, we maximize $\mathcal{L}$ with respect to μ_k and $\tilde{\pi}_k$. For μ_k, we have

$$\frac{\partial\mathcal{L}}{\partial\mu_k} = \sum_{n=1}^{N}\gamma(\tilde{z}_{nk})\left(\frac{y_n}{\mu_k} + \frac{1-y_n}{1-\mu_k}\right)$$

By the first order condition of the maximum, we have

$$\begin{aligned}
\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})\left(\frac{y_n}{\mu_k} - \frac{1-y_n}{1-\mu_k}\right) &= 0\\
\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})\frac{y_n-\mu_k}{\mu_k(1-\mu_k)} &= 0\\
\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})y_n - \sum_{n=1}^{N}\gamma(\tilde{z}_{nk})\mu_k &= 0\\
\mu_k &= \frac{\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})y_n}{\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})}
\end{aligned}$$

For $\tilde{\pi}_k$, we need a Lagrange multiplier to fulfil the constraint,

$$\tilde{\mathcal{L}} = \mathbb{E}_{\boldsymbol{z}}[\log p(\boldsymbol{y}, \tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{\pi}}, \boldsymbol{\mu})] + \lambda\left(\sum_{k=1}^{K}\tilde{\pi}_k - 1\right)$$

and

$$\frac{\partial\tilde{\mathcal{L}}}{\partial\tilde{\pi}_k} = \sum_{n=1}^{N}\gamma(\tilde{z}_{nk})\frac{1}{\tilde{\pi}_k} + \lambda$$

By the first order condition of the maximum, we have

$$\begin{aligned}
\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})\frac{1}{\tilde{\pi}_k} + \lambda &= 0\\
\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\tilde{z}_{nk}) + \sum_{k=1}^{K}\tilde{\pi}_k\lambda &= 0\\
\lambda = -\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\tilde{z}_{nk}) &= -N\\
\tilde{\pi}_k = -\frac{\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})}{\lambda} &= \frac{\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})}{N}
\end{aligned}$$

For our problem, we have $K = 2$. If $k = 1$, it means he wins the Switch game, $\tilde{\pi}_1 = \pi$, $\mu_1 = p$. If $k = 2$, it means he loses the Switch game, $\tilde{\pi}_2 = 1 - \pi$, $\mu_2 = q$.

E-step: Since $p = q$, all $\gamma(\tilde{z}_k)$ are the same,

$$\gamma(\tilde{z}_k) = p(z = 1 \mid y = 1) = \frac{p(z=1)p(y=1 \mid z=1)}{p(z=0)p(y=1 \mid z=0) + p(z=1)p(y=1 \mid z=1)} = \frac{\pi^{(0)}p^{(0)}}{\pi^{(0)}q^{(0)} + \pi^{(0)}p^{(0)}} = 0.5$$

M-step:

$$\pi^{(1)} = \frac{1}{N}\sum_{n=1}^{N} \gamma(\tilde{z}_k) = \frac{1}{10}(10)(0.5) = 0.5$$

$$p^{(1)} = \frac{\sum_{n=1}^{N} y_n \gamma(\tilde{z}_k)}{\sum_{n=1}^{N} \gamma(\tilde{z}_k)} = \frac{6(0.5)}{10(0.5)} = 0.6$$

$$q^{(1)} = \frac{\sum_{n=1}^{N} y_n \gamma(\tilde{z}_k)}{\sum_{n=1}^{N} \gamma(\tilde{z}_k)} = \frac{6(0.5)}{10(0.5)} = 0.6$$

■