CS 189 Introduction to Machine Learning
Spring 2018 HW9

Your self-grade URL is http://eecs189.org/self_grade?question_ids=1_1,1_2,2_1,2_2,3_1,3_2,3_3,4_1,4_2,4_3,4_4,4_5,4_6,5_1,5_2,5_3,5_4,6.

2 Classification Policy

Suppose we have a classification problem with classes labeled $1, \ldots, c$ and an additional "doubt" category labeled $c+1$. Let $f : \mathbb{R}^d \to \{1, \ldots, c+1\}$ be a decision rule. Define the loss function

$$L(f(\mathbf{x}), y) = \begin{cases} 0 & \text{if } f(\mathbf{x}) = y \quad f(\mathbf{x}) \in \{1, \ldots, c\}, \\ \lambda_c & \text{if } f(\mathbf{x}) \neq y \quad f(\mathbf{x}) \in \{1, \ldots, c\}, \\ \lambda_d & \text{if } f(\mathbf{x}) = c+1 \end{cases} \tag{1}$$

where $\lambda_c \geq 0$ is the loss incurred for making a misclassification and $\lambda_d \geq 0$ is the loss incurred for choosing doubt. In words this means the following:

- When you are correct, you should incur no loss.
- When you are incorrect, you should incur some penalty λ_c for making the wrong choice.
- When you are unsure about what to choose, you might want to select a category corresponding to "doubt" and you should incur a penalty λ_d.

We can see that in practice we'd like to have this sort of loss function if we don't want to make a decision if we are unsure about it. This sort of loss function, however, doesn't help you in instances where you have high certainty about a decision, but that decision is wrong.

To understand the expected amount of loss we will incur with decision rule $f(\mathbf{x})$, we look at the risk. The risk of classifying a new data point $\mathbf{x}$ as class $f(\mathbf{x}) \in \{1, 2, \ldots, c+1\}$ is

$$R(f(\mathbf{x})|\mathbf{x}) = \sum_{i=1}^{c} L(f(\mathbf{x}), i)\, P(Y = i|\mathbf{x}).$$

(a) **Show that the following policy $f_{opt}(x)$ obtains the minimum risk:**

- (**R1**) Find class i such that $P(Y = i|\mathbf{x}) \geq P(Y = j|\mathbf{x})$ for all j, meaning you pick the class with the highest probability given x.
- (**R2**) Choose class i if $P(Y = i|\mathbf{x}) \geq 1 - \frac{\lambda_d}{\lambda_c}$
- (**R3**) Choose doubt otherwise.

Solution:

- Let's first simplify the risk given our specific loss function. If $f(\mathbf{x}) = i$ where i is not doubt, then the risk is

$$R(f(\mathbf{x}) = i|\mathbf{x}) = \sum_{j=1}^{c} L(f(\mathbf{x}) = i, y = j)P(Y = j|\mathbf{x}) \tag{2}$$

$$= 0 \cdot P(Y = i|\mathbf{x}) + \lambda_c \sum_{j=1, j\neq i} P(Y = j|\mathbf{x}) \tag{3}$$

$$= \lambda_c \left(1 - P(Y = i|\mathbf{x})\right) \tag{4}$$

When $f(\mathbf{x}) = c + 1$, meaning you've chosen doubt, the risk is:

$$R(f(\mathbf{x}) = c + 1|\mathbf{x}) = \sum_{j=1}^{c} L(f(\mathbf{x}) = c + 1, y = j)P(Y = j|\mathbf{x}) \tag{5}$$

$$= \lambda_d \sum_{j=1} P(Y = j|\mathbf{x}) \tag{6}$$

$$= \lambda_d \tag{7}$$

because $\sum_{j=1} P(Y = j|\mathbf{x})$ should sum to 1 since its a proper probability distribution. Now let $f_{opt} : \mathbb{R}^d \to \{1, \ldots, c + 1\}$ be the decision rule which implements (**R1**)–(**R3**). We want to show that in expectation the rule f_{opt} is at least as good as an arbitrary rule f. Let $\mathbf{x} \in \mathbb{R}^d$ be a data point, which we want to classify. Let's examine all the possible scenarios where $f_{opt}(\mathbf{x})$ and another arbitrary rule $f(\mathbf{x})$ might differ:
Case 1: Let $f_{opt}(\mathbf{x}) = i$ where $i \neq c + 1$.

- Case 1a: $f(\mathbf{x}) = k$ where $k \neq i$. Then we get with (**R1**) that

$$\begin{aligned} R(f_{opt}(\mathbf{x}) = i|\mathbf{x}) &= \lambda_c \left(1 - P(Y = i|\mathbf{x})\right) \\ &\leq \lambda_c \left(1 - P(Y = k|\mathbf{x})\right) = R(f(\mathbf{x}) = k|\mathbf{x}). \end{aligned}$$

- Case 1b: $f(\mathbf{x}) = c + 1$. Then we get with (**R1**) that

$$\begin{aligned} R(f_{opt}(\mathbf{x}) = i|\mathbf{x}) &= \lambda_c \left(1 - P(Y = i|\mathbf{x})\right) \\ &\leq \lambda_c(1 - (1 - \frac{\lambda_d}{\lambda_c})) = \lambda_d = R(f(\mathbf{x}) = c + 1|\mathbf{x}). \end{aligned}$$

Case 2: Let $f_{opt}(\mathbf{x}) = c + 1$ and $f(\mathbf{x}) = k$ where $k \neq c + 1$. Then:

$$R(f(\mathbf{x}) = k|\mathbf{x}) = \lambda_c(1 - P(Y = k|\mathbf{x})$$

$$R(f_{opt}(\mathbf{x}) = c + 1|\mathbf{x}) = \lambda_d$$

We are in case (**R3**) which means that:

$$\max_{j\in\{1,\ldots,c\}} P(Y = j|\mathbf{x}) < 1 - \lambda_d/\lambda_c$$

hence $P(Y = k|\mathbf{x}) < 1 - \lambda_d/\lambda_c$, which means

$$R(f(\mathbf{x}) = k|\mathbf{x}) > \lambda_d = R(f_{opt}(\mathbf{x}))$$

Therefore in every case we proved that the rule f_{opt} is at least as good as the arbitrary rule f, which proves that f_{opt} is an optimal rule.

(b) **How would you modify your optimum decision rule if $\lambda_d = 0$? What happens if $\lambda_d > \lambda_c$? Explain why this is or is not consistent with what one would expect intuitively.**

Solution: If $\lambda_d = 0$, then property (1) will hold iff there exists an $i \in \{1, \ldots, c\}$ such that $P(f_{opt}(\mathbf{x}) = i|\mathbf{x}) = 1$. So we will either classify x in class i if we are 100% sure about this, or else we will choose doubt. Of course this is completely consistent with our intuition, because choosing doubt does not have any penalty at all, since $\lambda_d = 0$.

If $\lambda_d > \lambda_c$, then we will always classify x in the class $i \in \{1, \ldots, c\}$ which gives the highest probability of correct classification. Once again this makes sense, since the cost of choosing doubt is higher than classifying $\mathbf{x}$ in any of the classes, hence our best option is to classify x in the class which gives the highest probability for a correct classification.

3 LDA and CCA

Consider the following random variable $\mathbf{X} \in \mathbb{R}^d$, generated using a *mixture of two Gaussians*. Here, the vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^d$ are arbitrary (mean) vectors, and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ represents a positive definite (covariance) matrix. For now, we will assume that we know all of these parameters.

Draw a label $L \in \{1, 2\}$ such that the label 1 is chosen with probability π_1 (and consequently, label 2 with probability $\pi_2 = 1 - \pi_1$), and generate $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_L, \boldsymbol{\Sigma})$.

(a) Now given a particular $\mathbf{X} \in \mathbb{R}^d$ generated from the above model, we wish to find its label. **Write out the decision rule corresponding to the following estimates of L:**

- MLE (i.e. L is a parameter here that we want to estimate)
- MAP

Your decision rule should take the form of a threshold: if some function $f(\mathbf{X}) > T$, then choose the label 1, otherwise choose the label 2. **When are these two decision rules the same?** Hint: investigate the ratio between the two likelihood functions and the ratio between the two posterior probabilities respectively.

Solution: For finding the MLE, the two possible parameters values are $L = 1$ and $L = 2$ corresponding to probability distributions $p_1(\mathbf{x}) = P(\mathbf{X} = \mathbf{x} \mid L = 1)$ and $p_2(\mathbf{x}) = P(\mathbf{X} = \mathbf{x} \mid L = 2)$. To compute the MLE, we therefore need to find which of $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ is larger. Taking a ratio of the two, we have

$$\frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} = \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\}. \tag{8}$$

Let us now use the following simplifying fact. We have

$$\mathbf{v}^\top \mathbf{M}\mathbf{v} - \mathbf{u}^\top \mathbf{M}\mathbf{u} = \mathbf{v}^\top \mathbf{M}\mathbf{v} - \mathbf{v}^\top \mathbf{M}\mathbf{u} - (\mathbf{u} - \mathbf{v})^\top \mathbf{M}\mathbf{u}$$

$$= \mathbf{v}^\top \mathbf{M}(\mathbf{v}-\mathbf{u}) + (\mathbf{v}-\mathbf{u})^\top \mathbf{M}\mathbf{u}$$
$$= (\mathbf{v}+\mathbf{u})^\top \mathbf{M}(\mathbf{v}-\mathbf{u}),$$

where we have used the fact that $\mathbf{M}$ is symmetric.

Substituting $\mathbf{v} = \mathbf{x} - \boldsymbol{\mu}_2$, $\mathbf{u} = \mathbf{x} - \boldsymbol{\mu}_1$, we have

$$\log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} = \left(\mathbf{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Notice that we are interested in comparing this quantity with 0; if it is greater than zero, then we prefer 1 to 2, and we prefer 2 otherwise. The MLE decision rule therefore takes the form

$$\hat{L}_{\text{MLE}} = \begin{cases} 1 \text{ if } f(\mathbf{X}) = \left(\mathbf{X} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0 \\ 2 \text{ otherwise.} \end{cases}$$

Notice that this is clearly linear in $\mathbf{X}$.

For the MAP estimator, we are interested in comparing the quantities $P(L=1|\mathbf{X}=\mathbf{x})$ and $P(L=2|\mathbf{X}=\mathbf{x})$. By an application of Bayes rule, we are comparing $\frac{1}{P(\mathbf{X}=\mathbf{x})}P(\mathbf{X}=\mathbf{x}|L=1)P(L=1)$ with $\frac{1}{P(\mathbf{X}=\mathbf{x})}P(\mathbf{X}=\mathbf{x}|L=2)P(L=2)$, and so it suffices to compute the ratio

$$\frac{P(\mathbf{X}=\mathbf{x}|L=1)P(L=1)}{P(\mathbf{X}=\mathbf{x}|L=2)P(L=2)} = \frac{\pi_1 p_1(\mathbf{x})}{\pi_2 p_2(\mathbf{x})}.$$

Since we are going to compute the logarithm of this expression and compare it to zero, it suffices to check the condition

$$\log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} > \log \frac{\pi_2}{\pi_1}.$$

The thresholding is therefore identical to the calculation above up to an offset, and we have

$$\hat{L}_{\text{MAP}} = \begin{cases} 1 \text{ if } \left(\mathbf{X} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > \log \frac{\pi_2}{\pi_1} \\ 2 \text{ otherwise.} \end{cases}$$

Clearly, when $\pi_1 = \pi_2 = 1/2$, these two classifiers are exactly the same.

(b) You should have noticed that the classification function f is *linear* in its argument $\mathbf{X}$, and takes the form $\mathbf{w}^\top(\mathbf{X} - \mathbf{v})$. We will now show that CCA defined on a suitable set of random variables leads to precisely the same decision rule.

Let $\mathbf{Y} \in \mathbb{R}^2$ be a one hot vector denoting the realization of the label ℓ, i.e. if $L = \ell$ we have $Y_\ell = 1$ and otherwise $Y_\ell = 0$. Let $\pi_1 = \pi_2 = 1/2$. **Compute the covariance matrices $\boldsymbol{\Sigma}_{XX}, \boldsymbol{\Sigma}_{XY}$ and $\boldsymbol{\Sigma}_{YY}$ as a function of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$.** Recall that the random variables are not zero-mean. Hint: when computing the covariance matrices, the tower property of the expectation is useful.

Solution: Let us first compute the covariance matrix of $\mathbf{X}$. Recall that $\mathbf{X} \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ when $L = 1$ (which happens with probability $1/2$), and $\mathbf{X} \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ otherwise.

Also note that the covariance matrix of a non-zero mean RV $\mathbf{Z}$ with mean $\boldsymbol{\mu}$ can be written as $\mathbb{E}[(\mathbf{Z} - \boldsymbol{\mu})(\mathbf{Z} - \boldsymbol{\mu})^\top] = \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top] - 2\boldsymbol{\mu}\mathbb{E}[\mathbf{Z}^\top] + \boldsymbol{\mu}\boldsymbol{\mu}^\top = \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top$.

Let us now use $\boldsymbol{\mu} = \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}$ to denote the mean of $\mathbf{X}$. Hence, we have

$$\boldsymbol{\Sigma}_{XX} = \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top.$$

where

$$\begin{aligned}
\mathbb{E}[\mathbf{X}\mathbf{X}^\top] &= \frac{1}{2}\mathbb{E}[\mathbf{X}\mathbf{X}^\top | L = 1] + \frac{1}{2}\mathbb{E}[\mathbf{X}\mathbf{X}^\top | L = 2] \\
&= \frac{1}{2}\left(\boldsymbol{\Sigma} + \boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top\right) + \frac{1}{2}\left(\boldsymbol{\Sigma} + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top\right) \\
&= \boldsymbol{\Sigma} + \frac{1}{2}\left(\boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top\right).
\end{aligned}$$

Combining this with our calculation above, we have

$$\begin{aligned}
\boldsymbol{\Sigma}_{XX} &= \boldsymbol{\Sigma} + \frac{1}{2}\left(\boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top\right) - \boldsymbol{\mu}\boldsymbol{\mu}^\top \\
&= \boldsymbol{\Sigma} + \frac{1}{2}\left(\boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top\right) - \frac{1}{4}(\boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top + 2\boldsymbol{\mu}_1\boldsymbol{\mu}_2^\top \\
&= \boldsymbol{\Sigma} + \frac{1}{4}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top.
\end{aligned}$$

Let us now compute $\boldsymbol{\Sigma}_{YY}$. Using the same idea, and noting that $\mathbb{E}[\mathbf{Y}] = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$, we have

$$\boldsymbol{\Sigma}_{YY} = \mathbb{E}[\mathbf{Y}\mathbf{Y}^\top] - \begin{bmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{bmatrix}.$$

Also note that $\mathbb{E}[Y_i Y_j] = 0$ if $i \neq j$ (since one of the random variables will always be zero by the one-hot property), and $\mathbb{E}[Y_1^2] = \pi_1 = 1/2$ and $\mathbb{E}[Y_2^2] = \pi_2 = 1/2$.

Thus, the covariance matrix is given by

$$\begin{aligned}
\boldsymbol{\Sigma}_{YY} &= \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix} - \begin{bmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{bmatrix} \\
&= \begin{bmatrix} 1/4 & -1/4 \\ -1/4 & 1/4 \end{bmatrix} \\
&= \frac{1}{4}\begin{bmatrix} 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & -1 \end{bmatrix}^\top
\end{aligned}$$

To conclude, we must compute the covariance matrix Σ_{XY}. Note that we can again condition on the labels to accomplish this, and using $\mathbf{1/2}$ to denote the all half vector, we have

$$\boldsymbol{\Sigma}_{XY} = \frac{1}{2}\mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{Y} - \mathbf{1/2})^\top | L = 1] + \frac{1}{2}\mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{Y} - \mathbf{1/2})^\top | L = 2]$$

$$
\begin{aligned}
&= \frac{1}{2}\mathbb{E}[(\mathbf{X}-\boldsymbol{\mu})\begin{bmatrix}1/2 & -1/2\end{bmatrix}|L=1] + \frac{1}{2}\mathbb{E}[(\mathbf{X}-\boldsymbol{\mu})\begin{bmatrix}-1/2 & 1/2\end{bmatrix}|L=2] \\
&= \frac{1}{4}\mathbb{E}[\begin{bmatrix}(\mathbf{X}-\boldsymbol{\mu}) & -(\mathbf{X}-\boldsymbol{\mu})\end{bmatrix}|L=1] + \frac{1}{4}\mathbb{E}[\begin{bmatrix}-(\mathbf{X}-\boldsymbol{\mu}) & (\mathbf{X}-\boldsymbol{\mu})\end{bmatrix}|L=2] \\
&= \frac{1}{4}\begin{bmatrix}\frac{\boldsymbol{\mu}_1-\boldsymbol{\mu}_2}{2} & \frac{\boldsymbol{\mu}_2-\boldsymbol{\mu}_1}{2}\end{bmatrix} + \frac{1}{4}\begin{bmatrix}\frac{\boldsymbol{\mu}_1-\boldsymbol{\mu}_2}{2} & \frac{\boldsymbol{\mu}_2-\boldsymbol{\mu}_1}{2}\end{bmatrix} \\
&= \frac{1}{4}\begin{bmatrix}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2 & \boldsymbol{\mu}_2-\boldsymbol{\mu}_1\end{bmatrix} \\
&= \frac{1}{4}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\begin{bmatrix}1 & -1\end{bmatrix}.
\end{aligned}
$$

(c) Let us now perform CCA on the two random vectors $\mathbf{X}$ and $\mathbf{Y}$. Recall that in order to find the first canonical directions, we look for vectors $\mathbf{u} \in \mathbb{R}^d$ and $\mathbf{v} \in \mathbb{R}^2$ such that $\rho(\mathbf{u}^\top\mathbf{X}, \mathbf{v}^\top\mathbf{Y})$ is maximized.

Show that the maximizing $\mathbf{u}^*$ is proportional to $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Recall that $\mathbf{u}^*$ is that "direction" of $\mathbf{X}$ that contributes most to predicting $\mathbf{Y}$. **What is the relationship between $\mathbf{u}^*$ and the function $f(\mathbf{X})$ computed in part (a)?**

Hint: The Sherman-Morrison formula for matrix inversion may be useful:

Suppose $\mathbf{A} \in \mathbb{R}^{d\times d}$ is an invertible square matrix and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ are column vectors. Then,

$$(\mathbf{A} + \mathbf{a}\mathbf{b}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{a}\mathbf{b}^T\mathbf{A}^{-1}}{1 + \mathbf{b}^T\mathbf{A}^{-1}\mathbf{a}}.$$

Solution: Recall the definition of the correlation coefficient. We are interested in vectors $\mathbf{u}, \mathbf{v}$ such that we maximize

$$
\begin{aligned}
\rho(\mathbf{u}^\top\mathbf{X}, \mathbf{v}^\top\mathbf{Y}) &= \frac{\mathsf{cov}(\mathbf{u}^\top\mathbf{X}, \mathbf{v}^\top\mathbf{Y})}{\sqrt{\mathrm{Var}(\mathbf{u}^\top\mathbf{X})\,\mathrm{Var}(\mathbf{v}^\top\mathbf{Y})}} \\
&= \frac{\mathbf{u}^\top\boldsymbol{\Sigma}_{XY}\mathbf{v}}{\sqrt{\mathbf{u}^\top\boldsymbol{\Sigma}_{XX}\mathbf{u}}\sqrt{\mathbf{v}^\top\boldsymbol{\Sigma}_{YY}\mathbf{v}}}.
\end{aligned}
$$

Now using the expressions we computed above, we have

$$
\begin{aligned}
\mathbf{u}^\top\boldsymbol{\Sigma}_{XY}\mathbf{v} &= \frac{1}{4}\mathbf{u}^\top(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\begin{bmatrix}1 & -1\end{bmatrix}^\top\mathbf{v} \\
\mathbf{v}^\top\boldsymbol{\Sigma}_{YY}\mathbf{v} &= \frac{1}{4}\left(\begin{bmatrix}1 & -1\end{bmatrix}^\top\mathbf{v}\right)^2.
\end{aligned}
$$

Consequently, we see that the factor $\begin{bmatrix}1 & -1\end{bmatrix}^\top\mathbf{v}$ cancels in both the numerator and denominator, yielding

$$\rho(\mathbf{u}^\top\mathbf{X}, \mathbf{v}^\top\mathbf{Y}) = \frac{1}{2}\frac{\mathbf{u}^\top(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{\sqrt{\mathbf{u}^\top\boldsymbol{\Sigma}_{XX}\mathbf{u}}},$$

and we are interested in maximizing this expression over all $\mathbf{u}$. Performing the change of variables $\mathbf{w} = \mathbf{\Sigma}_{XX}^{1/2}\mathbf{u}$, we are interested in the expression

$$\max_{\mathbf{w}} \frac{\mathbf{w}^\top \mathbf{\Sigma}_{XX}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\|\mathbf{w}\|_2}.$$

Since the above expression does not change when $\mathbf{w}$ is multiplied by a constant, we may assume that it has unit Euclidean norm, and so the optimization problem becomes

$$\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top \mathbf{\Sigma}_{XX}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

To find the maximizer, we use Cauchy Schwarz inequality to obtain

$$\mathbf{w}^\top \mathbf{\Sigma}_{XX}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \leq \|\mathbf{w}\|_2 \|\mathbf{\Sigma}_{XX}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\|_2,$$

and equality is attained by the vector $\mathbf{w}^* = \alpha \mathbf{\Sigma}_{XX}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ for a scalar α that normalizes $\mathbf{w}^*$ to be unit norm.

We can then obtain $\mathbf{u}^* = \mathbf{\Sigma}_{XX}^{-1/2}\mathbf{w}^* = \alpha \mathbf{\Sigma}_{XX}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

We are now in a position to apply the hint. We know that

$$\mathbf{\Sigma}_{XX}^{-1} = \mathbf{\Sigma}^{-1} - \frac{\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1}}{4 + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)};$$

notice now that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is a scalar β. We also have $\mathbf{u}^* \propto \mathbf{\Sigma}_{XX}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{4+\beta}\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\beta$.

In other words, we have $\mathbf{u}^* = \gamma \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ for the component of $\mathbf{X}$ that has the highest predictive value for $\mathbf{Y}$. Comparing with the decision function $f(\mathbf{X}) = (\mathbf{X} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2})^\top \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, we see that this is the same direction predicted by LDA in part (a).

4 Sensors, Objects, and Localization (Part 2)

Let us say there are n objects and m sensors located in a $2d$ plane. The n objects are located at the points $(x_1, y_1), \ldots, (x_n, y_n)$. The m sensors are located at the points $(a_1, b_1), \ldots, (a_m, b_m)$. We have measurements for the distances between the objects and the sensors: D_{ij} is the measured distance from object i to sensor j. The distance measurement has noise in it. Specifically, we model

$$D_{ij} = ||(x_i, y_i) - (a_j, b_j)|| + Z_{ij},$$

where $Z_{ij} \sim N(0, 1)$. The noise is independent across different measurements.

Assume we observe $D_{ij} = d_{ij}$ with $(X_i, Y_i) = (x_i, y_i)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Here, $m = 7$. **Our goal is to predict $(X_{i'}, Y_{i'})$ from newly observed $D_{i'1}, \ldots, D_{i'7}$.** For a data set with q points, the error is measured by the average distance between the predicted object locations and the true object locations,

$$\frac{1}{q}\sum_{i=1}^{q} \sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2},$$

where $(\hat{x}_i, \hat{y}_i)$ is the location of objects predicted by a model.

We are going to consider six models in this problem and compare their performance:

- A *Generative Model*: This is basically from the earlier assignment: we first estimate sensor locations from the training data by solving the nonlinear learst squares problem using the Gauss-Newton algorithm[1], and then use the estimated sensor locations to estimate the new object locations.
- An *Oracle Model*: This is the same as generative model except that we will use the ground truth sensor location rather than the estimated sensor location.
- A *Linear Model*. Using the training set, the linear model attempts to fit (X_i, Y_i) directly from the distance measurements $(D_{i1}, \ldots, D_{i7})$. Then it predicts $(X_{i'}, Y_{i'})$ from $(D_{i'1}, \ldots, D_{i'7})$ using the map that it found during training. (It never tries to explicitly model the underlying sensor locations.)
- A *Second-Order Polynomial Regression Model*. The set-up is similar to the linear model, but including second-order polynomial features.
- A *Third-Order Polynomial Regression Model*. The set-up is similar to the linear model, but including third-order polynomial features.
- A *Neural Network Model*. The Neural Network should have two hidden layers, each with 100 neurons, and use ReLU and/or tanh for the non-linearity. (You are encouraged to explore on your own beyond this however. These parameters were chosen to teach you a hype-deflating lesson.) The neural net approach also follows the principle of finding/learning a direct connection between the distance measurements and the object location.

(a) **Implement the last four models listed above in `models.py`.** Starter code has been provided for data generation and visualization to aid your explorations. We provide you a simple gradient descent framework for you to implement the neural network, but you are also free to use the TensorFlow and PyTorch code from your previous homework and other 3rd libraries.

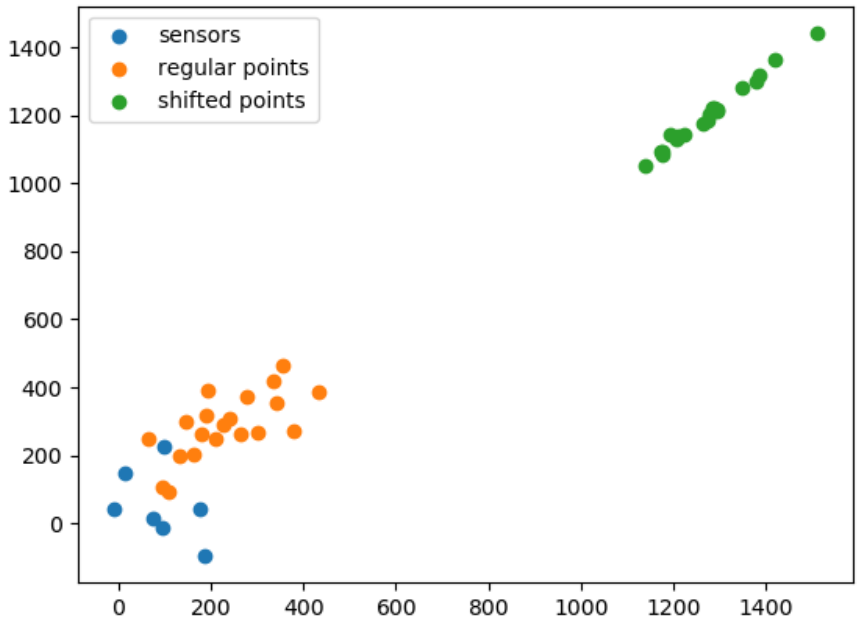
Solution: See the code that we released.

(b) In the following parts, we will deal with two types of data, the "regular" data set, and the "shifted" data set. The "regular" data set has the same distribution as the training data set, while the "shifted" data set has a different distribution. **Run `plot0.py` to visualize** the sensor location, the distribution of the "regular" data set, and the distribution of the "shifted" data set. **Attach the plot**.

Solution:

See the attached figure.

[1]This is not covered in the class but you can find the related information in `https://www.wikiwand.com/en/Gauss-Newton_algorithm`



(c) The starter code generated a set of 7 sensors and the following data sets:

- 15 training sets where n_{train} varies from 10 to 290 in increments of 20.
- A "regular" testing data set where $n_{\text{test}} = 1000$.
- A "shifted" testing data set where $n_{\text{test}} = 1000$. You can do this by setting original_dist to `False` in the function `generate_data` in `starter.py`.

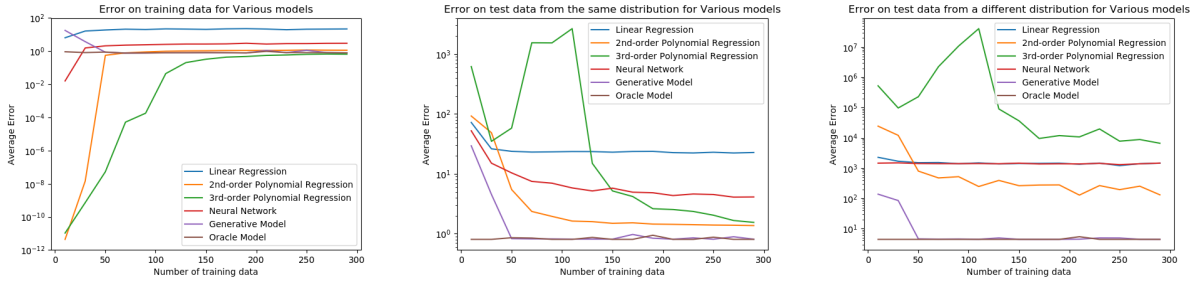
Use `plot1.py` to train each of the five models on each of the fifteen training sets. Use your results to generate three figures. Each figure should include *all* of the models on the same plot so that you can compare them:

- A plot of *training error* versus n_{train} (the amount of data used to train the model) for all of the models.
- A plot of *testing error* on the "regular" test set versus n_{train} (the amount of data used to train the model) for all of the models.
- A plot of *testing error* on the "shifted" test set versus n_{train} (the amount of data used to train the model) for all of the models.

Briefly describe your observations. What are the strengths and weaknesses of each model?

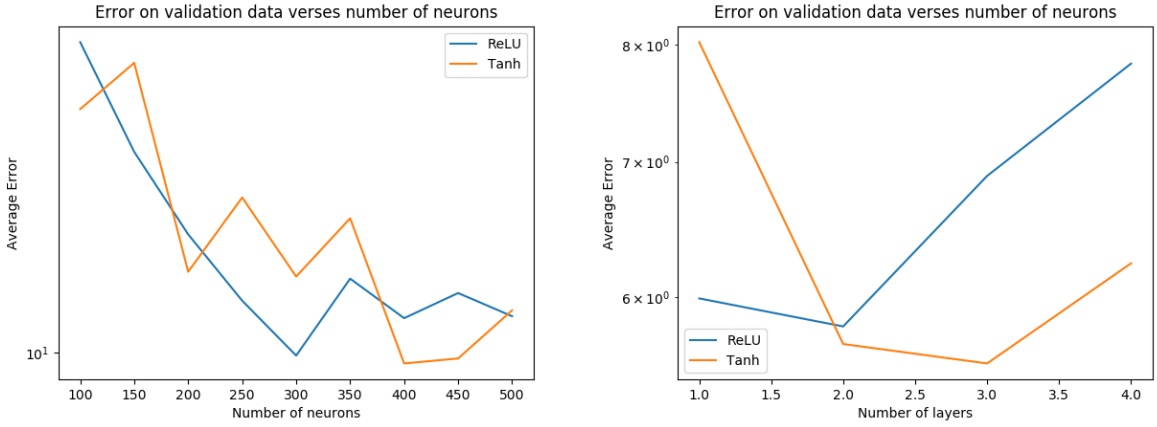
Solution: The plot by varying the numbers of data points n_{train} from 10 to 290 by 20 is shown. We observe that the generative model achieves the best performance as expected. Why? When data is drawn from a different distribution, we see that although all the models are performing worse, the generative model is still generalizing at some reasonable level. The rest of

the models are behaving quite poorly indeed. This is because extrapolation is fundamentally challenging for any universal-function-approximation type model. Second-order polynomial has the second better performance. Third order polynomial performs well when the data is generated from the same distribution, followed by neural networks. The linear model fails to perform well even with ample training data.



(d) We are now going to do some hyper-parameter tuning on our neural network. Fix the number of hidden layers to be two and let ℓ be the number of neurons in each of these two layers. Try values for ℓ between 100 and 500 in increments of 50. Use data sets with $n_{\text{train}} = 200$ and $n_{\text{test}} = 1,000$. **What is the best choice for ℓ (the number of neurons in the hidden layers)? Justify your answer with plots.** The starter code is in `plot2.py`.

Solution: The best performance is achieved at the case when number of neurons equals to $300 - 500$. (Any reasonable answer that is supported with a plot is acceptable.)



(e) We are going to do some more hyper-parameter tuning on our neural network. Let k be the number of hidden layers and let ℓ be the number of neurons in each hidden layer. **Write a formula for the total number of weights in our network in terms of ℓ and k. If we want to keep the total number of weights in the network approximately equal to 10000, find a formula for ℓ in terms of k.** Try values of k between 1 and 4 with the appropriate implied choice for ℓ. Use data sets with $n_{\text{train}} = 200$ and $n_{\text{test}} = 1000$. **What is the best choice for k (the number of layers)? Justify your answer with plots.** The starter code is in `plot3.py`.

Solution: For $k > 1$: The number of neurons n_{neurons} for given k and ℓ, it is enough to use the approximate formula $n_{\text{neurons}} \approx (k-1)\ell^2$ (each interior layer has ℓ^2 neurons). We therefore get $\ell \approx 100/\sqrt{k-1}$. We get the following numbers $k = 1$: $\ell \approx 1100$, $k = 2$: $\ell \approx 100$, $k = 3$: $\ell \approx 70$ and $k = 4$: $\ell \approx 50$.

The best performance is achieved at the case when number of layers equals to 3 with tanh and 2 with ReLU. (Any reasonable answer that is supported with a plot is acceptable.)

(f) You might have seen that the neural network performance is disappointing compared to the generative model in the "shifted" data. Try increasing the number of training data and tune the hyper-parameters. Can you get it to generalize to the "shifted" test data? **Attach the "number of training data vs accuracy" plot to justify your conclusion.** What is the intuition how the neural network works on predicting D? The starter kit is provided in `plot4.py`.

Solution:

Figure 1: Generalization of Neural Network

No, increasing the training data does not work well as we saw in Figure 1. The reason is that the data are generated from a different distribution and neural network cannot transfer its performance to that, as mentioned above. Performance of the tuned neural network on data generated from a different distribution can be seen in the figures above. The neural network works like using a nonlinear metric to interpolate the test data using the training samples. When the distribution of testing data is not similar to the training data, it is hard for the neural network to figure it out without the prior knowledge of the underlying model.

5 Entropy, KL Divergences, and Cross-Entropy

Stepping back for a bit, notice that so far we have mostly considered so called *Euclidean* spaces, the most prominent example is the vector space $\mathbb{R}^d$ with inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$ and norm

$\|\mathbf{x}\|_2^2 = \mathbf{x}^\top \mathbf{x}$. For example in linear regression, we had the loss function

$$f(\boldsymbol{\theta}) = \|\mathbf{X}^\top \boldsymbol{\theta} - \mathbf{y}\|_2^2 = \sum_{i=1}^{n} (\mathbf{x}_i^\top \boldsymbol{\theta} - y_i)^2$$

which uses precisely the Euclidean norm squared to measure the error.

In this problem we will implicitly consider a different geometric structure, which defines a metric on probability distributions. In more advanced courses, this can be developed further to show how probability distributions are naturally imbued with a curved non-Euclidean intrinsic geometry. Here, our goals are more modest — we just want you to better understand the relationship between probability distributions, entropy, KL Divergence, and cross-entropy.

Let $\mathbf{p}$ and $\mathbf{q}$ be two probability distributions, i.e. $p_i \geq 0$, $q_i \geq 0$, $\sum_i p_i = 1$ and $\sum_i q_i = 1$, then we define the Kullback-Leibler divergence

$$\mathrm{KL}(\mathbf{p}, \mathbf{q}) = \sum_i p_i \log \frac{p_i}{q_i}$$

which is the "distance" to $\mathbf{p}$ from $\mathbf{q}$. We have $\mathrm{KL}(\mathbf{p}, \mathbf{p}) = 0$ and $\mathrm{KL}(\mathbf{p}, \mathbf{q}) \geq 0$ (the latter by Jensen's inequality) as would be expected from a distance metric. However, $\mathrm{KL}(\mathbf{p}, \mathbf{q}) \neq \mathrm{KL}(\mathbf{q}, \mathbf{p})$ since the KL divergence is not symmetric.

(a) *Entropy motivation:* Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables taking values in a finite set $\{0, 1, \ldots, m\}$, i.e. $p_j = P(X_i = j)$ for $j \in \{0, 1, \ldots, m\}$. The *empirical number of occurrences* is then a random vector that we can denote $\mathbf{F}^{(n)}$ where $F_j^{(n)}$ is the number of variables X_i that happen to take a value equal to j.

Intuitively, we can consider coin tosses with $j = 0$ corresponding to heads and $j = 1$ corresponding to tails. Say we do an experiment with $n = 100$ coin tosses, then $F_0^{(100)}$ is the number of heads that came up and $F_1^{(100)}$ is the number of tails.

Recall that the number of configurations of $X_1, X_2, \ldots, X_n$ that have $f^{(n)}$ as their empirical type is $\binom{n}{f_0^{(n)}, f_1^{(n)}, \ldots, f_m^{(n)}}$. Further notice that dividing the empirical type by n yeilds an empirical probability distribution.

Show using the crudest form of Stirling's approximation ($\ell! \approx (\frac{\ell}{e})^\ell$) that this is approximately equal to $\exp\Big(nH(f^{(n)}/n)\Big)$ where the entropy H of a probability distribution is defined as $H(\mathbf{p}) = \sum_{j=0}^{m} p_j \ln \frac{1}{p_j}$.

Solution: Note: The multinomial coefficient

$$\binom{n}{f_0^{(n)}, f_1^{(n)}, \ldots, f_m^{(n)}} = \frac{n!}{f_0^{(n)}! f_1^{(n)}! \cdots f_m^{(n)}!}$$

is the number of ways to put n interchangeable objects into m boxes so that box j has $f_j^{(n)}$ objects in it. There are that many distinct "n-length strings" of X realizations whose empirical counts match $\mathbf{f}^{(n)}$.

By applying the Stirling approximation, we get

$$\begin{aligned}\binom{n}{f_0^{(n)}, f_1^{(n)}, \ldots, f_m^{(n)}} &= \frac{n!}{f_0^{(n)}! f_1^{(n)}! \cdots f_m^{(n)}!} \\ &\approx \frac{(n/e)^n}{(f_0^{(n)}/e)^{f_0^{(n)}} (f_1^{(n)}/e)^{f_1^{(n)}} \cdots (f_m^{(n)}/e)^{f_m^{(n)}}} \\ &= \prod_{i=0}^{m} \left(\frac{n}{f_i^{(n)}}\right)^{f_i^{(n)}} = \exp\left(n \sum_{i=0}^{m} \frac{f_i^{(n)}}{n} \log\left(\frac{n}{f_i^{(n)}}\right)\right) \\ &= \exp\left(nH(f^{(n)}/n)\right).\end{aligned}$$

(b) *KL divergence motivation:* Recall that the probability of seeing a particular empirical type is given by:

$$P(\mathbf{F}^{(n)} = \mathbf{f}^{(n)}) = \binom{n}{f_0^{(n)}, f_1^{(n)}, \ldots, f_m^{(n)}} \prod_{j=1}^{m} p_j^{f_j^{(n)}}.$$

Consider the limit of large n and a sequence of empirical types so that $\frac{1}{n}\mathbf{f}^{(n)} \to \mathbf{f}$ for $n \to \infty$, where $\mathbf{f}$ is some distribution of interest.

Use Stirling's approximation to show that

$$\lim_{n\to\infty} \frac{1}{n} \log P(\mathbf{F}^{(n)} = \mathbf{f}^{(n)}) = -\operatorname{KL}(\mathbf{f}, \mathbf{p})$$

Intuitively this means that the larger $\operatorname{KL}(\mathbf{f}, \mathbf{p})$ is, the easier it is to conclude $\mathbf{f} \neq \mathbf{p}$ from empirical data since the chance that we would get confused in that way is decaying exponentially. Note also that the empirical distribution is the first argument of the KL divergence and the true model is the second argument of the KL divergence — we are going from the true model to the empirical one.

Solution: Recall that there are

$$\binom{n}{f_0^{(n)}, f_1^{(n)}, \ldots, f_m^{(n)}} = \frac{n!}{f_0^{(n)}! f_1^{(n)}! \cdots f_m^{(n)}!}$$

many ways to realize $\mathbf{f}^{(n)}$, each of them has probability $\prod_{j=1}^{m} p_j^{f_j^{(n)}}$. Therefore

$$P(\mathbf{F}^{(n)} = \mathbf{f}^{(n)}) = \binom{n}{f_0^{(n)}, f_1^{(n)}, \ldots, f_m^{(n)}} \prod_{j=1}^{m} p_j^{f_j^{(n)}}.$$

Using the alternate form of Stirling's formula $\log(n!) \approx n \log n - n + o(n)$ for large n we then get

$$\log P(F^{(n)} = f^{(n)}) = \log \frac{n!}{f_0^{(n)}! f_1^{(n)}! \cdots f_m^{(n)}!} \prod_{j=1}^{m} p_j^{f_j^{(n)}}$$

$$\approx (n \log n - n) - \sum_j (f_j^{(n)} \log f_j^{(n)} - f_j^{(n)}) + \sum_j f_j^{(n)} \log p_j$$

$$= (n \log n - \sum_j f_j^{(n)} \log n) - (n - \sum_j f_j^{(n)}) - \sum_j f_j^{(n)} \log \frac{f_j^{(n)}}{n} + \sum_j f_j^{(n)} \log p_j$$

$$= n \sum_j \frac{f_j^{(n)}}{n} \log \frac{p_j}{f_j^{(n)}/n}$$

Therefore we have

$$\frac{1}{n} \log P(F^{(n)} = f^{(n)}) \to \sum_j \frac{f_j^{(n)}}{n} \log \frac{p_j}{f_j^{(n)}/n} \to -\operatorname{KL}(f, p)$$

for $n \to \infty$.

(c) **Show that for probability distributions $p(\mathbf{x}, y)$ and $q_\theta(\mathbf{x}, y) = q_\theta(y \mid \mathbf{x})q(\mathbf{x})$ with $\mathbf{x}$ from some discrete set $\mathcal{X}$ and y from some discrete set $\mathcal{Y}$ we have**

$$\operatorname{KL}(p, q_\theta) = c - \sum_{\mathbf{x} \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(\mathbf{x}, y) \log q_\theta(y \mid \mathbf{x}) \tag{9}$$

for some constant c independent of θ. Solution: We have

$$\begin{aligned}
\operatorname{KL}(p, q_\theta) &= \sum_{\mathbf{x} \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(\mathbf{x}, y) \log \frac{p(\mathbf{x}, y)}{q_\theta(\mathbf{x}, y)} \\
&= \sum_{\mathbf{x} \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Big(p(\mathbf{x}, y) \log p(\mathbf{x}, y) - p(\mathbf{x}, y) \log \big(q_\theta(y \mid \mathbf{x}) q(\mathbf{x}) \big) \Big) \\
&= c - \sum_{\mathbf{x} \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(\mathbf{x}, y) \log q_\theta(y \mid \mathbf{x})
\end{aligned}$$

(d) In logistic regression we predict labels $y_i = +1$ or $y_i = -1$ from features $\mathbf{x}_i$ using the transition probability model

$$q_\theta(y_i \mid \mathbf{x}_i) = \frac{1}{1 + e^{-y_i \boldsymbol{\theta}^\top \mathbf{x}_i}}. \tag{10}$$

We now show that the cross-entropy loss you have seen in lectures can be formulated as minimizing the KL distance to the empirical probabilities from the probabilities induced by the model q_θ.

For convenience, we assume that all the feature $\mathbf{x}_i$ are distinct — no two training points are identical.

Use (c) to show that with the empirical distribution

$$p(\mathbf{x}, y) = \begin{cases} \frac{1}{n} & \text{if } \mathbf{x} = \mathbf{x}_i \text{ and } y = y_i \text{ for some } i = 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

we get

$$\min_{\boldsymbol{\theta}} \mathrm{KL}(p, q_{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}} -\frac{1}{n} \sum_i \log q_{\boldsymbol{\theta}}(y_i \mid \mathbf{x}_i),$$

which is the cross entropy loss derived in lectures.

Solution: By plugging $p(\mathbf{x}, y) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{\mathbf{x}=\mathbf{x}_i, y=y_i\}}$ into the expression from (c), we get

$$\begin{aligned} \mathrm{KL}(p, q_\theta) &= c - \sum_{\mathbf{x} \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^{n} 1_{\{\mathbf{x}=\mathbf{x}_i, y=y_i\}} \log q_\theta(y \mid \mathbf{x}) \\ &= c - \frac{1}{n} \sum_{i=1}^{n} \log q_\theta(y_i \mid \mathbf{x}_i). \end{aligned}$$

6 Your Own Question

Write your own question, and provide a thorough solution.

Writing your own problems is a very important way to really learn the material. The famous "Bloom's Taxonomy" that lists the levels of learning is: Remember, Understand, Apply, Analyze, Evaluate, and Create. Using what you know to create is the top-level. We rarely ask you any HW questions about the lowest level of straight-up remembering, expecting you to be able to do that yourself. (e.g. make yourself flashcards) But we don't want the same to be true about the highest level.

As a practical matter, having some practice at trying to create problems helps you study for exams much better than simply counting on solving existing practice problems. This is because thinking about how to create an interesting problem forces you to really look at the material from the perspective of those who are going to create the exams.

Besides, this is fun. If you want to make a boring problem, go ahead. That is your prerogative. But it is more fun to really engage with the material, discover something interesting, and then come up with a problem that walks others down a journey that lets them share your discovery. You don't have to achieve this every week. But unless you try every week, it probably won't happen ever.