ESE 504-542 : Statistics for Data Science
Instructor: Hamed Hassani
Fall 2019

Final Examination

NAME	

One two-sided note-sheet allowed.

	Grade (y/n)	Score	Max. Score
Problem 1			40
Problem 2			30
Problem 3			30
TOTAL			100

Problem 1 (40 points)

Recall that in classification we assume that each data point is an i.i.d. sample from a distribution $P(X = x, Y = y)$. In this question, we are going to consider a specific data distribution P and evaluate the performance of logistic regression and Bayes optimal classifier on data generated using P. In the following, we assume $x \in \mathbb{R}$ and $y \in \{-1, 1\}$, i.e. the data is one-dimensional and the label is binary. Write $P(X = x, Y = y) = P(Y = y)P(X = x|Y = y)$. We let $P(y = +1) = P(Y = -1) = \frac{1}{2}$ and

$$P(X = x|Y = +1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-5)^2}{2}), \text{ and,}$$

$$P(X = x|Y = -1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x+5)^2}{2}).$$

1. Start from $P(X = x, Y = y) = P(Y = y)P(X = x|Y = y)$ and show that $P(X = x, Y = y) = \frac{1}{2\sqrt{2\pi}} \exp(-\frac{(x-5y)^2}{2})$. (This is a simple one line derivation.)

We can re-state given equations into:

$$P(X=x \mid Y=y) = \frac{1}{\sqrt{2\pi}} exp(-\frac{(X-5y)^2}{2})$$

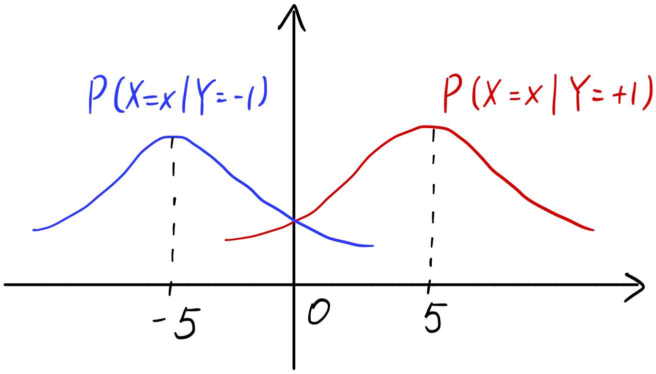
$$\therefore \; P(X=x, Y=y)$$

$$= P(Y=y)\; P(X=x|Y=y)$$

$$= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} exp(-\frac{(X-5y)^2}{2})$$

$$= \frac{1}{2\sqrt{2\pi}} exp(-\frac{(X-5y)^2}{2})$$

2. Plot the conditional distributions $P(X = x|Y = +1)$ and $P(X = x|Y = -1)$ in one figure. I.e. you should plot two gaussian pdfs in one figure.



3. Write the Bayes optimal classification rule given the above distribution P and simplify it (hint: in the end you should reach to a very simple classification rule that classifies an input x based on whether or not its value is greater than a threshold).

$$P(Y=y \mid X=x) = \frac{P(X=x, Y=y)}{P(X=x)}$$

$$= P(Y=y \mid X=x) \cdot \frac{P(Y=y)}{P(X=x)}$$

$$\therefore \hat{y} = h(x) = \underset{y}{\operatorname{argmax}} \; P(Y=y \mid X=x)$$

$$= \underset{y}{\operatorname{argmax}} \; P(Y=y \mid X=x) \cdot \frac{P(Y=y)}{P(X=x)}$$

$$= \underset{y}{\operatorname{argmax}} \; P(X=x \mid Y=y)$$

$$= \begin{cases} +1 & , \text{if } x > 0 \\ -1 & , \text{otherwise} \end{cases} \quad \text{by plot}$$

4. Compute the probability of classification error for the Bayes optimal classifier.

error

$$= \mathbb{E}_{(X,Y)\sim P}\left[\mathbb{1}(h(x) \neq Y) \right]$$

$$= P_{(X,Y)\sim P}\left(h(x) \neq Y \right)$$

$$= P(X>0)\, P(Y=-1 \mid X=x) + P(X<0)\, P(Y=+1 \mid X<0)$$

$$= P(Y=-1, X>0) + P(Y=+1, X<0)$$

$$= P(Y=-1)\, P(X>0 \mid Y=-1) + P(Y=+1)\, P(X<0 \mid Y=+1)$$

$$= 2 \cdot \frac{1}{2}\left(1 - \Phi\left(\frac{5}{1}\right)\right)$$

$$= 1 - \Phi(5)$$

5. Let us now consider logistic regression (this part and the next can be answered independently from the previous parts). Given training data $(x_1, y_1), \cdots, (x_n, y_n)$, explain briefly the main steps of training a logistic regression model. I.e. what quantities/probabilities are being estimated by logistic regression? What is the parametric model used? How are the parameters of the model optimized?

Estimate: $P(Y=+1|X=x)$, $P(Y=-1|X=x)$

Model: $P(Y=+1|X=x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$

$$P(Y=-1|X=x) = \frac{1}{1+e^{\beta_0+\beta_1 x}}$$

Optimization:

$$L(\beta_0, \beta_1|x) = Pr(Y|X; \beta_0, \beta_1)$$

$$= \prod_{i=1}^{n} P(Y=y_i|X=x_i; \beta_0, \beta_1)$$

$$\max_{\beta_0, \beta_1} L(\beta_0, \beta_1|x)$$

6. Going back to the data distribution P detailed above, logistic regression needs to find the value of two parameters β_0 and β_1 using training data $\{(x_i, y_i)\}_{i=1,\cdots,n}$ generated according to the distribution P. Assume that the number of training data points is very large (i.e. $n \to \infty$); What will be the parameters β_0 and β_1 in this case? (Hint: Start by deriving the exact form of the conditional distribution $P(Y = y|X = x)$.)

From : $P(X=x \mid Y=+1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-5)^2}{2}),\ P(Y=+1)=\frac{1}{2}$

$P(X=x \mid Y=-1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x+5)^2}{2}),\ P(Y=-1)=\frac{1}{2}$

$$
\begin{aligned}
& P(Y=+1 \mid X=x) \\
= & \frac{P(X=x,\ Y=+1)}{P(X=x,\ Y=+1)+P(X=x,\ Y=-1)} \\
= & \frac{P(X=x \mid Y=+1)P(Y=+1)}{P(X=x \mid Y=+1)P(Y=+1)+P(X=x \mid Y=-1)P(Y=-1)} \\
= & \frac{\frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-5)^2}{2})}{\frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-5)^2}{2})+\frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x+5)^2}{2})} \\
= & \frac{\exp(-\frac{(x-5)^2}{2})}{\exp(-\frac{(x-5)^2}{2})+\exp(-\frac{(x+5)^2}{2})} \\
= & \frac{1}{\exp(\frac{(x-5)^2}{2}-\frac{(x+5)^2}{2})+1} \\
= & \frac{1}{\exp(-10x)+1}
\end{aligned}
$$

Note that logistic regression would try estimate

$P(Y=+1 \mid X=x) = \frac{1}{1+\exp(-\beta_1 x-\beta_0)}$

When n is large, which means there is enough data to estimate probability accurately,

So $\beta_0 = 0,\ \beta_1 = 10.$

Problem 2 (30 points) [**Weighted K-Means Clustering**.]
Consider data points $x_1, x_2, \cdots, x_n \in \mathbb{R}^p$. We aim to provide an algorithm for the following clustering problem: Find K centers $c_1, c_2, \cdots, c_K \in \mathbb{R}^p$ that minimize the objective

$$\sum_{i=1}^{n} \min_{j \in \{1,\cdots,K\}} ||x_i - c_j||_1, \tag{1}$$

where $|| \cdot ||_1$ is the L_1 (Manhattan) distance.

1. Assume that $K = 1$. Find the optimal centroid that minimizes (1).

$K = 1$

$$f(c) = \sum_{i=1}^{n} || X_i - c ||_1$$

$$= \sum_{i=1}^{n} \sum_{\ell=1}^{p} | X_{i\ell} - C_\ell |$$

$$\frac{\partial f}{\partial c} = \sum_{i=1}^{n} \sum_{\ell=1}^{p} \text{sign} (C_\ell - X_{i\ell}) = 0$$

$\therefore \; |\{ x_i | X_{i\ell} > C_\ell \}| = |\{ x_i | X_{i\ell} < C_\ell \}|$

$\therefore$ C_ℓ is the median of

$X_{1\ell}, X_{2\ell} \cdots X_{n\ell}$, for all $1 \leq \ell \leq p$

2. For a given K, derive an iterative algorithm to find a good set of centers for the above problem. Explain precisely what the algorithm is and justify your answer. (Hint: Recall the steps of the K-Means algorithm done in class and see how you could change those for the above setting).

1. Randomly initialize K centers

2. Assign data points to cluster with nearest (least Manhattan distance) center.

3. Re-calculate centers with C_{jl} to be the median of $\{X_{jl} \mid X_j \in C_j\}$, for all $1 \leq l \leq p$

4. Repeat 2, 3 until centers don't update.

Problem 3 (30 points) **[Short answer questions]**

1. In which cases should we prefer LDA to QDA and vice versa? Briefly justify your answer.

LDA requires less parameters, and
is more likely to underfit and
less likely to overfit than QDA.
So we prefer LDA to QDA
when dataset is small
(we don't want overfitting)
and vice versa when dataset is large
(we don't want underfitting)

2. Assume that we have a data matrix X of dimension $p \times n$ as usual. Suppose that its SVD is of the from $X = USV^T$, where S is a diagonal matrix with $s_1 = 10$ and $s_2 = s_3 = \cdots = s_p = 1$. Assume that we want to compress the data from p to 1 dimensions via a linear transform represented by a $1 \times p$ matrix C and then reconstruct via $p \times 1$ matrix R. Let $\tilde{X} = RCX$ be the reconstruction. What is the smallest reconstruction error that can be achieved?

Compress data from p to 1 dimension :

$U = [u_1, \cdots, u_p]$, $C = u_1^T$

We choose the first principal component with $R = u_1$, since the largest eigenvalue leads to the smallest reconstruction error.

$$\|X - \tilde{X}\|_2^2 = \|X - RCX\|_2^2 = \|USV^T - US^{(1)}V^T\|_2^2$$

$$S^{(1)}_{ij} = 10 \cdot 1 \,(i=1,\ j=1)$$

$$\therefore \|X - \tilde{X}\|_2^2 = \|U(S - S^{(1)})V^T\|_2^2$$

$$= \|S - S^{(1)}\|_2^2$$

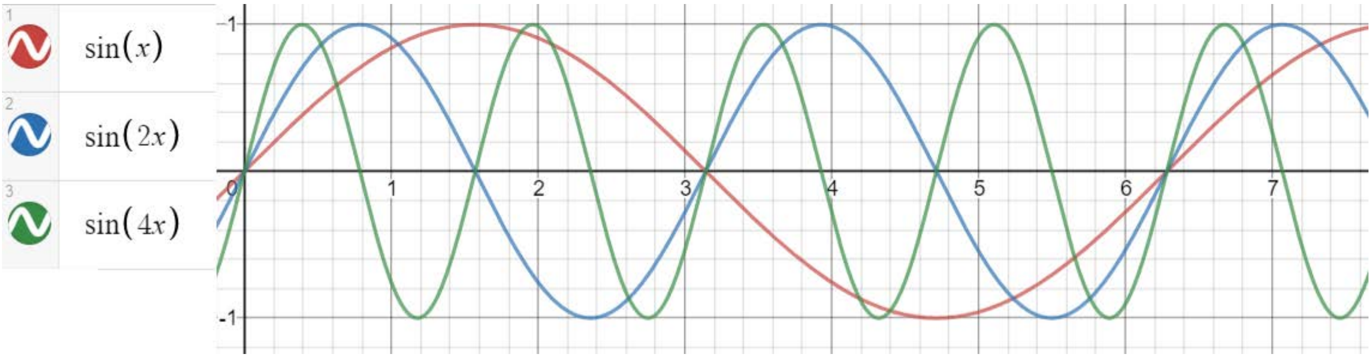
$$= \sum_{i=2}^{p} s_i^2$$

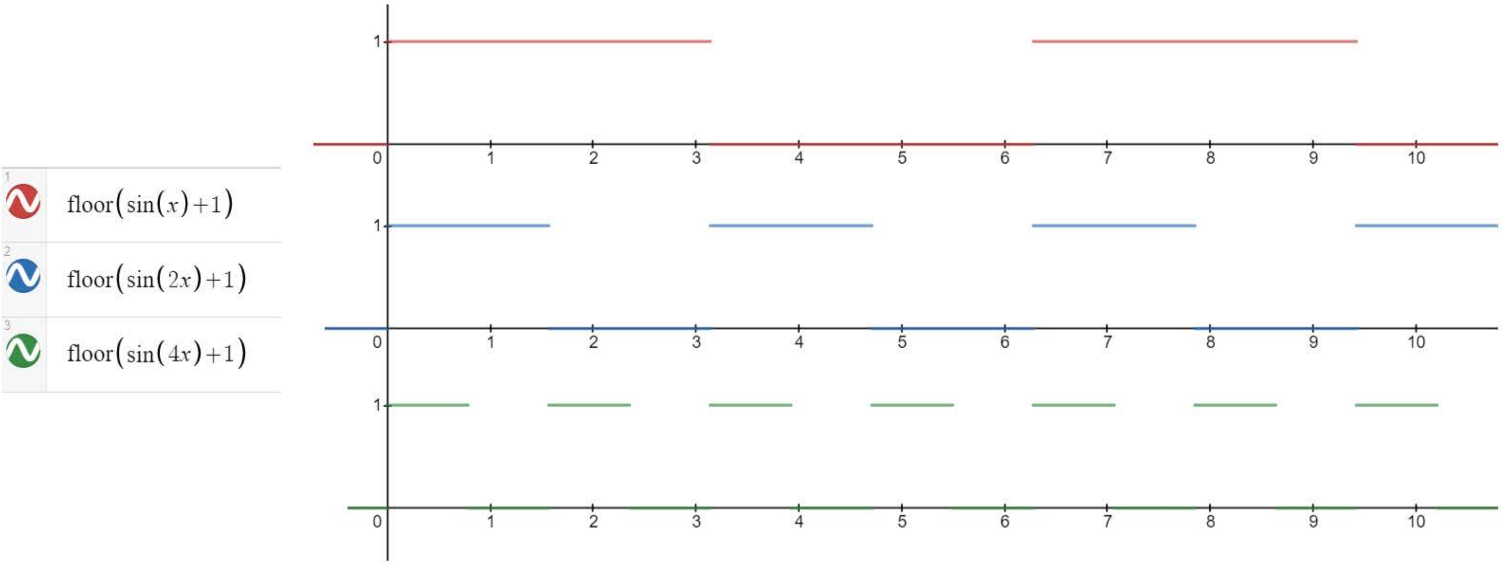
$$= p - 1$$

(If we choose other principle components, $\|X - \tilde{X}\|_2^2 = p + 8 > p - 1$)

3. Consider the hypothesis class $\mathcal{H} = \{h_\theta(x) = \lceil \sin(\theta x) \rceil, \theta \in \mathbb{R}\}$. What is the VC dimension of $\mathcal{H}$? Recall that $\lceil y \rceil$ is the smallest integer that's larger than y. (hint: Try to plot a few functions from $\mathcal{H}$ for different values of θ.)

Intuitively, when θ increases, $\sin(\theta x)$ become more and more complex. Below is illustration :





So $\mathcal{H}$ should shatter sets of any cardinality, VC-dim $= \infty$.

And following (not required) is a way to construct such set C.