Midterm Exam

Midterm exam will be done individually. Use of partial or entire solutions obtained from others or online is strictly prohibited.

- There will be 7 pages in this exam (including this cover sheet)
- This is a **CLOSED-BOOK** exam. You can use your one-sided half-page cheatsheet. You **CANNOT** use materials brought by other students.
- If you need more room to work out your answer to a question, use the back of the page and clearly mark this on the front of the page.
- Work efficiently and independently.
- You have 150 minutes.
- Good luck!

Question	Topic	Max. score	Score
1	Bayesian Decision Theory	15	
2	Logistic Regression	15	
3	Perceptron Algorithm	15	
4	Nonparametric Method	15	
5	Maximum Likelihood Estimator	20	
6	Short Answer Questions	20	
Total		100	

1. **Bayesian Decision Theory** (15 points)

   (a) (5 pts) Assume we have c classes $\omega_1, \ldots, \omega_c$, and feature vector $\mathbf{x}$, write down the Bayesian decision rule for classification in terms of prior probabilities of classes, i.e., $P(\omega_i)$, and class conditional densities of $\mathbf{x}$ i.e., $p(\mathbf{x}|\omega_i)$.

   (b) (10 pt) Consider two-class classification, suppose $p(\mathbf{x}|\omega_1)$ is standard normal distribution and $p(\mathbf{x}|\omega_2)$ is uniform distribution over $[-1, 1]$, i.e.,

$$\begin{aligned} p(\mathbf{x}|\omega_1) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\ p(\mathbf{x}|\omega_2) &= \frac{1}{2},\ \mathbf{x} \in [-1, 1] \end{aligned}$$

   Assuming zero-one loss and $P(\omega_1) = P(\omega_2)$, derive the corresponding decision rule.

2. **Logistic Regression** (15 points) In the lecture, we consider the logistic regression for binary classification on N observations $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$:

$$f(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x})$$

where $y_i \in \{0, 1\}$, $\sigma(.)$ is sigmoid function, and denote $\mathbf{y} = (y_1, y_2, \ldots, y_N)^T$. We showed to learn $\mathbf{w}$, maximizing the likelihood $p(\mathbf{y}|\mathbf{w})$ is equivalent to minimize the following *cross entropy* error function:

$$E(\mathbf{w}) = \sum_{i=1}^{N} -\{y_i \log f(\mathbf{x}_i) + (1 - y_i) \log(1 - f(\mathbf{x}_i))\}$$

Now consider the penalized version with the following error function:

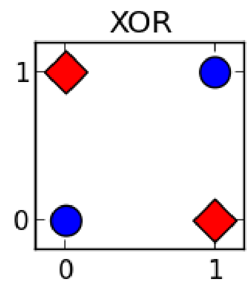
$$E_p(\mathbf{w}) = \sum_{i=1}^{N} -\{y_i \log f(\mathbf{x}_i) + (1 - y_i) \log(1 - f(\mathbf{x}_i))\} + \frac{\lambda}{2}||\mathbf{w}||^2$$

(a) (7 pts) Write down the gradient of $E_p(\mathbf{w})$ w.r.t $\mathbf{w}$, i.e., $\nabla_{\mathbf{w}} E_p(\mathbf{w})$.

(b) (8 pts) In terms of probabilities, minimizing the above penalized error function $E_p(\mathbf{w})$ is equivalent to maximizing the corresponding posterior $p(\mathbf{w}|\mathbf{y})$. Show that it is indeed the case by writing down the likelihood $p(\mathbf{y}|\mathbf{w})$ and prior $p(\mathbf{w})$.

3. **Perceptron Algorithm** (15 points) We consider the linear discriminant function for boolean function $f(x_1, x_2)$. x_1 and x_2 can be either 0 (false) or 1 (true). Think of boolean function as having 4 points on the 2D plane (x_1 being the horizontal axis and x_2 being the vertical axis).

   (a) (5 pts) For XOR problem, where we defined as $f(0,0) =$ **false**, $f(1,0) =$ **true**, $f(0,1) =$ **true**, and $f(1,1) =$ **false** where **true** can be treated as *positive class (red marker)* and **false** can be treated as *negative class (blue marker)*. Can we use perceptron algorithm to solve XOR problem? Why or why not?



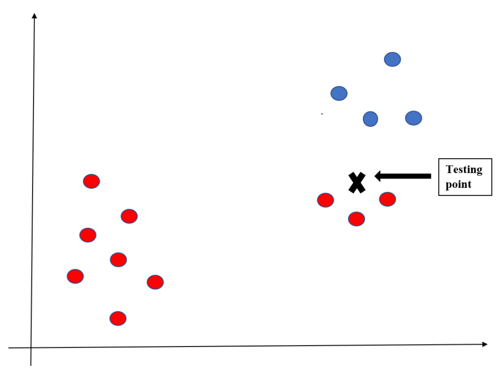
   (b) (10 pts) Now, consider OR function, where we defined as $f(0,0) =$ **false**, $f(1,0) =$ **true**, $f(0,1) =$ **true**, and $f(1,1) =$ **true** where **true** can be treated as *positive class* and **false** can be treated as *negative class*. Use perceptron algorithm to derive a linear decision boundary, please write down the steps. Suppose the initial decision boundary is $x_2 = \frac{1}{2}$, and you could start from the top-left point and sweep the whole data points in clockwise manner.

4. **Nonparametric Methods** (15 points)

(a) (8 pts)[1D Parzen window] Suppose we have 5 points $\{0, 1, 2, 4, 5\}$ that coming from 1D unknown distribution $p(x)$. Use Parzen window to estimate density value at $x = 3$. Assume we use *unit interval* as our window function $k(u)$, and the window width (interval length) $h = 3$:

$$k(u) = \begin{cases} 1, & |u| \leq 1/2 \\ 0, & \text{otherwise} \end{cases}$$

(b) (7 pts)[K nearest neighbour] Suppose we have the following data points that comes from two classes (red vs blue), X is the testing point. Distance is measured using Euclidean distance. Consider K nearest neighbour approach for testing point classification.



i. (2 pts) If $K = 1$, what is the class of X?

ii. (2 pts) If $K = 7$, what is the class of X?

iii. (3 pts) In general, use large value of K might be a good choice. Based on our data points, would you recommend use $K = 9$ nearest neighbours? Why or why not?

5. **Maximum Likelihood Estimator** (20 points) Suppose we have training samples $\{x_1, x_2, \ldots, x_n\}$. Consider the following distributions: (please write down the derivation steps)

   (a) (10 pts) $f(x;\theta) = \theta e^{-\theta x}$, $x \geq 0$, $\theta > 0$, find MLE for θ

   (b) (10 pts) $Unif[\theta_1, \theta_2]$, $Unif$ represents uniform distribution, find the MLE for θ_1 and θ_2.

6. **Short Answer Questions** (20 points)

(a) (5 pts) Describe main differences between generative approach, discriminative approach and discriminant function approach for classification.

(b) (5 pts) What are two criteria for Fisher's Linear Discriminant?

(c) (5 pts) Briefly compare the Principal Component Analysis (PCA) and Fisher's Linear Discriminant (FLD).

(d) (5 pts) Describe the main difference between batch gradient descent, mini-batch gradient descent and stochastic gradient descent. What are the pros and cons for batch gradient descent?