CSCI 5521: Introduction to Machine Learning (Spring 2024)[1]

Midterm Exam

Due on Gradescope by 01:00 pm, Mar 22nd

Instructions:

- This test has 4 questions, 100+2 points, including one extra credit problem worth 2 points.
- Please **write your name & ID on your submission pages**.
- **For full credit, show how you arrive at your answers**.
- You have **24 hours** to complete and submit this test to gradescope.

1. (**30 points**) In I-III, fill in the correct option(s) in the following table (it is not necessary to explain).

(I)	(II)	(III)

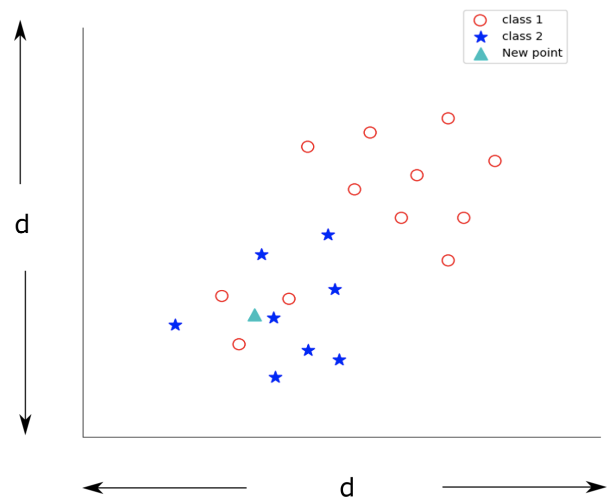
I. Select all the option(s) that correspond to supervised-learning algorithms:
(a) Principal component analysis
(b) Linear discriminant analysis
(c) k-means for clustering
(d) Nonparametric classification with a kernel estimator
(e) Linear discrimination

II. Which of the following option(s) help reduce overfitting in classification?
(a) Adding training data when performing classification
(b) Adding test data when performing classification
(c) Performing dimensionality reduction on all data before running a classifier
(d) Reducing the number of the parameters in the classifier
(e) Increasing the number of categories (e.g., from binary classification with $K = 2$ to multi-class classification with $K > 2$) when performing classification

III. Select all the true statement(s) below:
(a) In the training stage of an unsupervised classification task, the model takes in unlabeled data and outputs the model.
(b) In the testing stage of a supervised classification task, the model takes in unlabeled data and outputs the label.
(c) Principled component analysis and linear discriminant analysis are different methods for dimensionality reduction, and therefore must suggest different dimensions for projection.
(d) An objective function is always one to be minimized.
(e) Both gradient descent and EM find global optimum.

[1]Instructor: Catherine Qi Zhao. TA: Xianyu Chen, Amarachi Nzeukwu, James Yang, Yifeng Zhang. Email: csci5521.s2024@gmail.com

2. (**24 points**) Given a set of data points $\{x^t\}$ each shown in the figure, find the label of a new data point x using different non-parametric estimators / classifications as specified below.



(a) Write down the label of the new data point x with k nearest neighbor estimator when $k = 10$. Briefly explain the reason.

(b) Write down the label of the new data point x with k nearest neighbor estimator when $k = 20$. Briefly explain the reason.

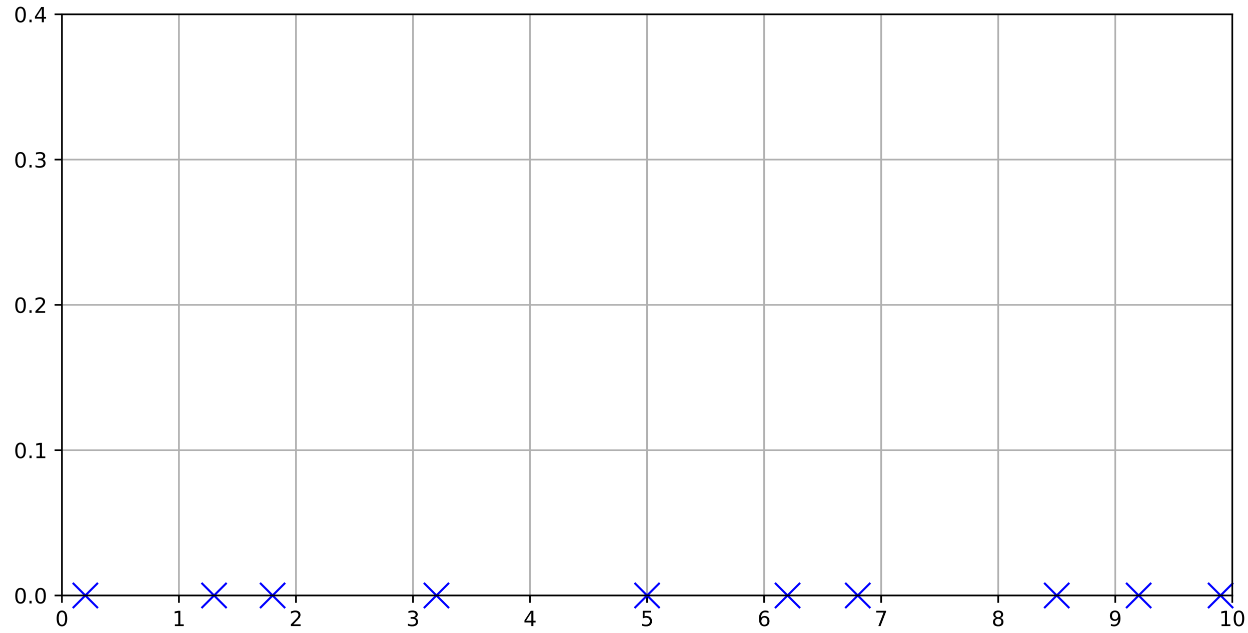
(c) Assume a uniform kernel function:

$$K(x, x^t) = \begin{cases} \frac{1}{\pi d^2}, & ||x - x^t||_2 \le d \\ 0, & \text{otherwise} \end{cases}$$

Write down the label of the new data point x with kernel estimator. Briefly explain the reason.

(d) (Extra credit, 2 points) Analyze the case when we use a kernel estimator with a Gaussian kernel (i.e., analyze the changes with the label with respect to different parameters of the Gaussian).

3. (**26 points**) Answer the following questions about nonparametric density estimator:



(a) Draw a histogram estimator (start from origin) using $h = 2$ for the following 10 training data points in $\mathbb{R}$: 0.2, 1.3, 1.8, 3.2, 5.0, 6.2, 6.8, 8.5, 9.2, 9.9

(b) Given a test data point $x = 5.5$, what is the predicted density $p(x)$ for the data point?

(c) List one possible approach to get a smoother density estimate.

(d) Draw an approximate curve when the kernel is used. Discuss the difference with and without kernel used. You do not need to show the calculation.

4. (**20 points**) In binary classification, gradient descent is used to find the minimum of the error function. With error function defined as the binary cross entropy loss $E(w, w_0|X) = -\sum_t \big(r^t \log y^t + (1-r^t)\log(1-y^t)\big)$ and $y^t = tanh(w^T x^t + w_0)$, **derive the update for w_j and w_0, where $j \neq 0$.**

   Note: Given $y = tanh(\alpha)$, the derivative $\frac{\partial y}{\partial \alpha} = 1 - y^2$.