Name+Email ID Jon-Michael Hoang | hoang339

CSCI 5521: Machine Learning Fundamentals (Spring 2022)[1]

Midterm Exam

Due on Gradescope by 1pm, Mar 18

Instructions:

- This test has 4 questions, 100 points + 3 points (extra credits), on 4 pages.
- **For full credit, show how you arrive at your answers**.
- You have **24 hours** to complete and submit this test to gradescope.

1. (**32 points**) In I-IV, select the correct option(s) (it is not necessary to explain).

I. Select all the option(s) that correspond to supervised-learning algorithms:
(a) k-Means
(b) Linear discriminant analysis
(c) Linear discrimination
(d) Linear regression

II. What could we do to reduce overfitting? Select all the option(s) that apply:
(a) Add more training data and keep testing data the same
(b) Add more testing data and keep training data the same
(c) Change to a less complex model (e.g., a model with less parameters)
(d) Use mixture of Gaussians instead of k-means for clustering

Fits less accurately but fast.

Fits more accurately towards the data, but costly

III. Select all the option(s) that correspond to dimensionality reduction algorithms:
(a) Linear discriminant analysis (b) Mixture of Gaussians
(c) Principal component analysis (d) Expectation maximization

IV. Select all the true statement(s) below.:
(a) Both MLE and MAP are parametric.
(b) Both MLE and MAP consider prior information.
(c) Expectation Maximization finds local optimum instead of global ones.
(d) Gradient descent finds global optimum instead of local ones.

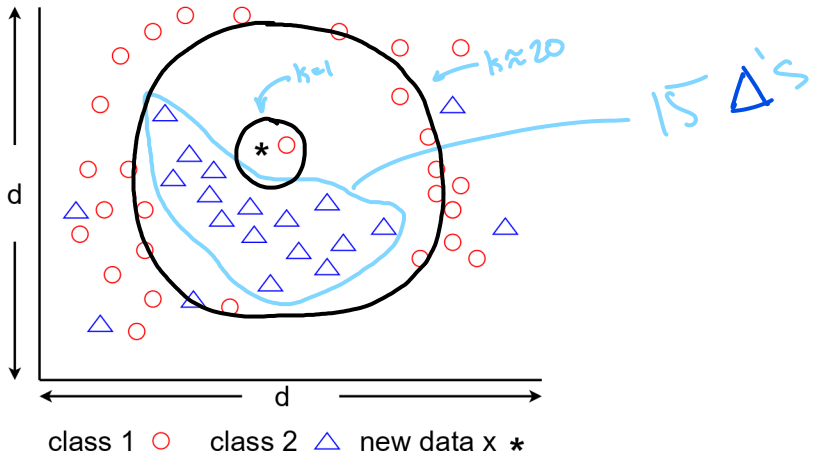
Lecture 5

priors are assumed

priors aren't assumed b/c uniform

[1]Instructor: Catherine Qi Zhao. TA: Prithvi Raj Botcha, Shi Chen, Suzie Hoops, James Yang, Yifeng Zhang. Email: csci5521.s2022@gmail.com

2. (**22 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 $k = 1$. Briefly explain the reason. $x \in$ ○ b/c the 1st nearest neighbor is ○

(b) Write down the label of the new data point x with k nearest neighbor estimator when $k = 20$. Briefly explain the reason. $x \in \Delta$ b/c the first 20 nearest neighbors are majorly Δ (~15)

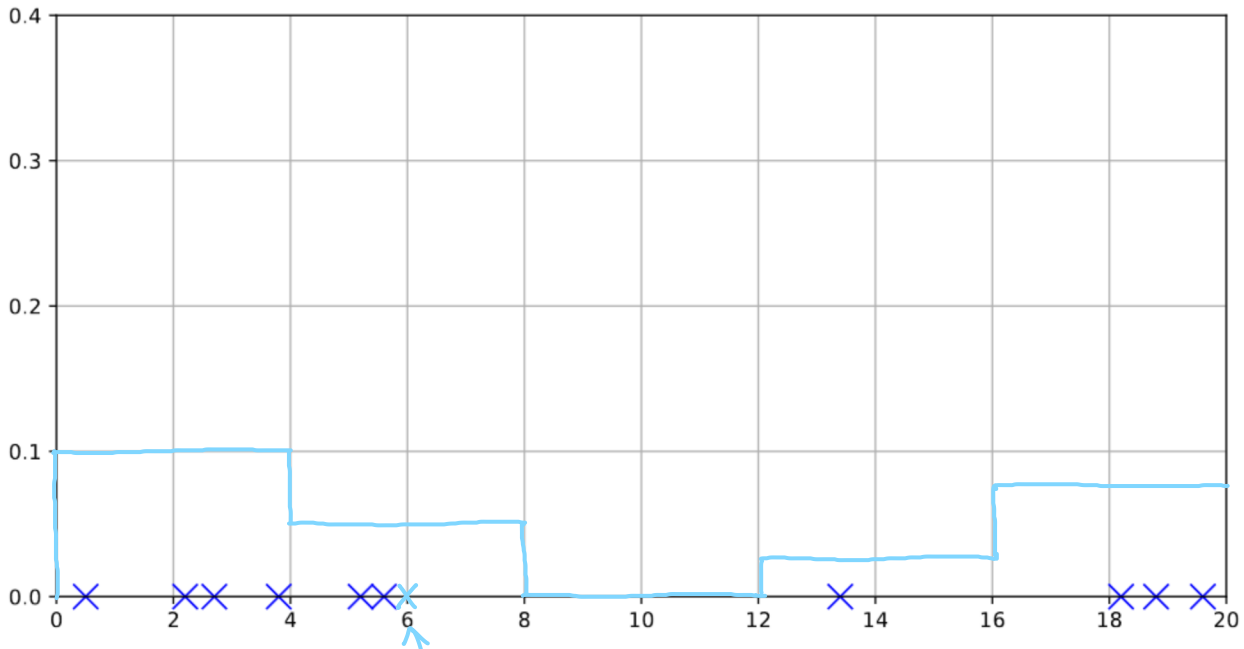
(c) Assume a uniform kernel function:

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

Write down the label of the new data point x with kernel estimator. Justify your answer.

The kernel takes into account all samples in the graph based on distance d.
Each point shares the same weight, thus the prediction is based on the number of samples for each class.
∴ $x \in$ ○ b/c there are more ○'s than Δ's in the d-space.

3. (**24 points**) Answer the following questions about non-parametric density estimator:



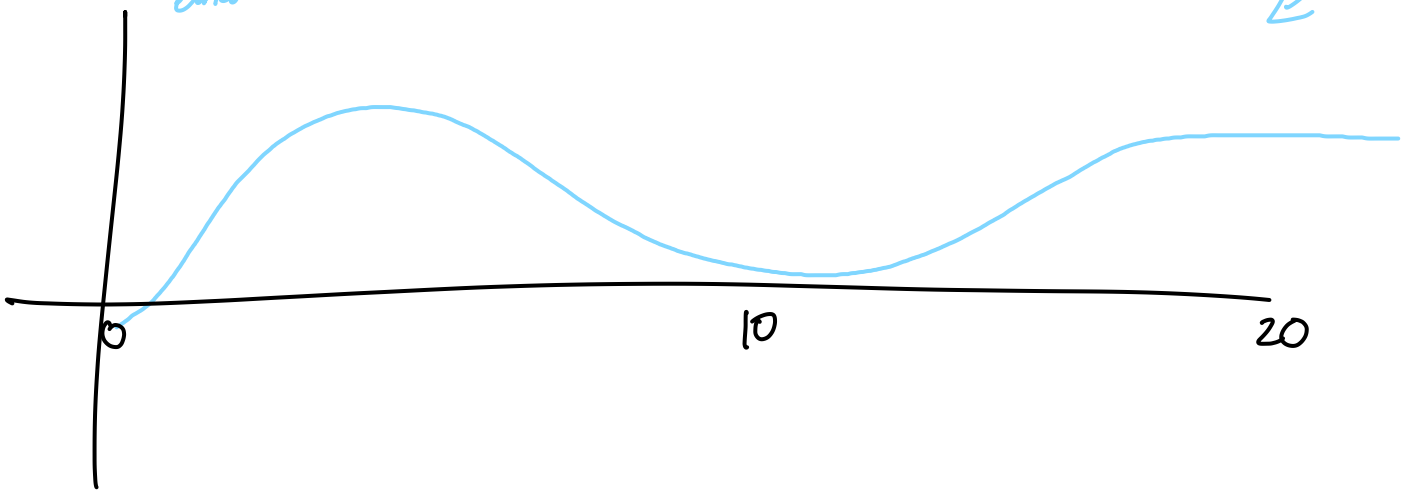
(a) Draw a histogram estimator (start from origin) using $h = 4.0$ for the following 10 training data points in $\mathbb{R}$: 0.5, 2.2, 2.7, 3.8, 5.2, 5.6, 13.4, 18.2, 18.8, 19.6 $\frac{\#\text{in bin}}{\text{Total in } \mathbb{R} * h} = \frac{\#\text{in bin}}{10 * 4} = \frac{\#\text{in bin}}{40}$

(b) Given a test data point $x = 6.0$, what is the predicted density $p(x)$ for the data point? Show your work for full credits.

$p(x=6.0) = 0.05$

(c) List one possible approach to get a smoother density estimate. Draw an approximate curve when the kernel is used. You do not need to show the calculation.

One possible approach to get a smoother density estimate is to use a kernel on the histogram wherein we can interpolate b/t the two nearest bin centers. We can consider the bin centers as x^t, histogram values as r^t, and use any interpolation scheme, linear or kernel based.



4. (**22+3 points**) Given a set of N iid random variables $\mathcal{X} = x_1, ...x_N$ and $\mathcal{Z} \in \mathbb{R}^{N \times K}$, the likelihood function is given by:

$$\mathcal{L}(\pi, \theta | \mathcal{X}, \mathcal{Z}) = p(\mathcal{X}, \mathcal{Z} | \pi, \theta) = log \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} Bern(x_n | \theta_k)^{z_{nk}} \tag{1}$$

such that $\sum_{k=1}^{K} \pi_k = 1$, $\sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} = N$ and $Bern(x|\theta_k) = \theta_k^x (1-\theta_k)^{1-x}$.

(a) Derive the MLE of θ_k.

(b) (extra credits, 3 points) Derive the MLE of π_k. Hint: there is a constraint on π_k.

* log → ln, for ease.

$\Rightarrow \mathcal{L}(\pi, \theta | \mathcal{X}, \mathcal{Z}) = p(\mathcal{X}, \mathcal{Z} | \pi, \theta) = \ln \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} Bern(x_n|\theta_k)^{z_{nk}}$

a) $\Rightarrow \sum_n \sum_k \ln\left(\pi_k^{z_{nk}} \left(\theta_k^{x_n} (1-\theta_k)^{1-x_n}\right)^{z_{nk}}\right) \Rightarrow \sum_n \sum_k \left(z_{nk} \ln \pi_k + z_{nk} \underbrace{\ln\left(\theta_k^{x_n} (1-\theta_k)^{1-x_n}\right)}_{x_n \ln \theta_k + (1-x_n)\ln(1-\theta_k)}\right)$

$\Rightarrow \sum_n \sum_k (z_{nk} \ln \pi_k) + \sum_n \sum_k (z_{nk}(x_n \ln \theta_k + (1-x_n)\ln(1-\theta_k)))$

$\Rightarrow \frac{\partial}{\partial \theta} = \sum_n \sum_k \left(\frac{z_{nk}(\theta_k - x_n)}{\theta_k^2 - \theta_k}\right) = 0$

$\Rightarrow \frac{N(\theta_k - x_n)}{\theta^2 - \theta_k} = \frac{0}{N} \Rightarrow \boxed{\hat{\theta}_k = x_1}$

$\frac{N}{N}$

b) $\frac{\partial}{\partial \pi} = \sum_n \sum_k \frac{z_{nk}}{\pi_k} \Rightarrow \sum_n \frac{1}{\pi_k} \sum_k z_{nk} = 0$

$= \frac{N}{\pi_k} = 0 \Rightarrow \boxed{\hat{\pi}_k = \frac{1}{\sqrt{0}}}$