STAT 451 Exam2

Summer 2024

1. Consider using k-means on the unsupervised 1D dataset $\{\mathbf{x}\} = \{0, 3, 6, 7, 8\}$ to create $k = 2$ clusters. Suppose the two initial randomly-chosen cluster centroids are $\mathbf{c}_1 = 6$ and $\mathbf{c}_2 = 7$.

   (a) What are the centroids after the first iteration of k-means?

   $\mathbf{c}_1 =$ [ ] and $\mathbf{c}_2 =$ [ ].

   (b) What are the centroids after the second iteration?

   $\mathbf{c}_1 =$ [ ] and $\mathbf{c}_2 =$ [ ].

2. For each situation, indicate which hyperparameter search strategy, grid search or random search, is more likely to be successful. Suppose computation time is limited.

   (a) A model has two hyperparameters:

   - The first is from the normal distribution with mean 6 and standard deviation 2.
   - The second is an integer in the range $[-10000, 10000]$.

   ◯ Grid search
   ◯ Random search

   (b) A model has three hyperparameters:

   - The first takes a floating-point number from the set $\{1, e, \pi\}$.
   - The second is a metric, either 'euclidean', 'manhattan', or a user-defined function.
   - The third is the number of CPU cores available, from 1 to 3.

   ◯ Grid search
   ◯ Random search

3. Mark each statement true or false.

(a) If $0 < d < p < D$, then using PCA to transform D-dimensional data $\{\mathbf{x}_i\}$ to p-dimensional data and then using PCA again to transform those p-dimensional data to d-dimensional data gives the same transformed data as using PCA to transform the D-dimensional data directly to d-dimensional data.

◯ True
◯ False

(b) An SVM makes a classification error on $(\mathbf{x}, y = -1)$ when $y(\mathbf{wx} + b) > 0$.

◯ True
◯ False

(c) In logistic regression, we predict $\hat{y} = 1$ for $\mathbf{x}$ if $\frac{1}{1+e^{-(\mathbf{wx}+b)}} > 0$.

◯ True
◯ False

(d) We would get the same model in linear regression if we chose parameters to minimize the logarithm of MSE instead of choosing them to minimize MSE itself.

◯ True
◯ False

(e) When splitting a binary decision tree node S containing $\{y\} = \{0, 0, 1, 1, 1\}$, the tree will prefer the split $S_- = \{0, 0, 1, 1\}$ and $S_+ = \{1\}$ to the split $S_- = \{0, 0, 1\}$ and $S_+ = \{1, 1\}$.

◯ True
◯ False

(f) In a linear support vector machine, $\mathbf{w}$ is normal to all three lines given by $\mathbf{wx} + b = -1$, $\mathbf{wx} + b = 0$, and $\mathbf{wx} + b = 1$.

◯ True
◯ False

(g) Ridge regression can address the problem of ordinary least squares regression overfitting the training data.

◯ True
◯ False

(h) If we run k-means clustering with $k = 2$ on the 1D data $\{0, 2, 3, 4, 5\}$ starting with cluster centers 0 and 5, the final centers are 2 and 4.

◯ True
◯ False

(i) If we run DBSCAN with $\epsilon = 2$ and $n = 3$ on the 1D data $\{0, 2, 3, 4\}$ starting at 3, the final clusters are $\{0\}$ and $\{2, 3, 4\}$.

◯ True
◯ False

(j) The bagging and random forest methods both use resampling with replacement from the training examples.

- ◯ True
- ◯ False

(k) Boosting selects a different random subset of features to process on each iteration.

- ◯ True
- ◯ False

(l) k-NN classification may benefit from standardizing features.

- ◯ True
- ◯ False

(m) Decision tree regression may benefit from standardizing features.

- ◯ True
- ◯ False

(n) Simple linear regression (in which $\mathbf{x}$ is a 1D scalar x) may benefit from standardizing the feature x.

- ◯ True
- ◯ False

(o) In PCA, the first principal component is the one in the direction of the greatest variability in the data.

- ◯ True
- ◯ False

(p) $1 + 1 = 2$.

- ◯ True
- ◯ False

4. Which of the following methods are typically used for feature selection or feature extraction? Choose Yes or No for each method.

(a) Lasso regression

- ◯ Yes.
- ◯ No.

(b) Kernel trick

- ◯ Yes.
- ◯ No.

(c) Correlation

- ◯ Yes.
- ◯ No.

(d) Principal component analysis

- ◯ Yes.
- ◯ No.

(e) Kernel density estimation

- ◯ Yes.
- ◯ No.

(f) Gradient descent

- ◯ Yes.
- ◯ No.

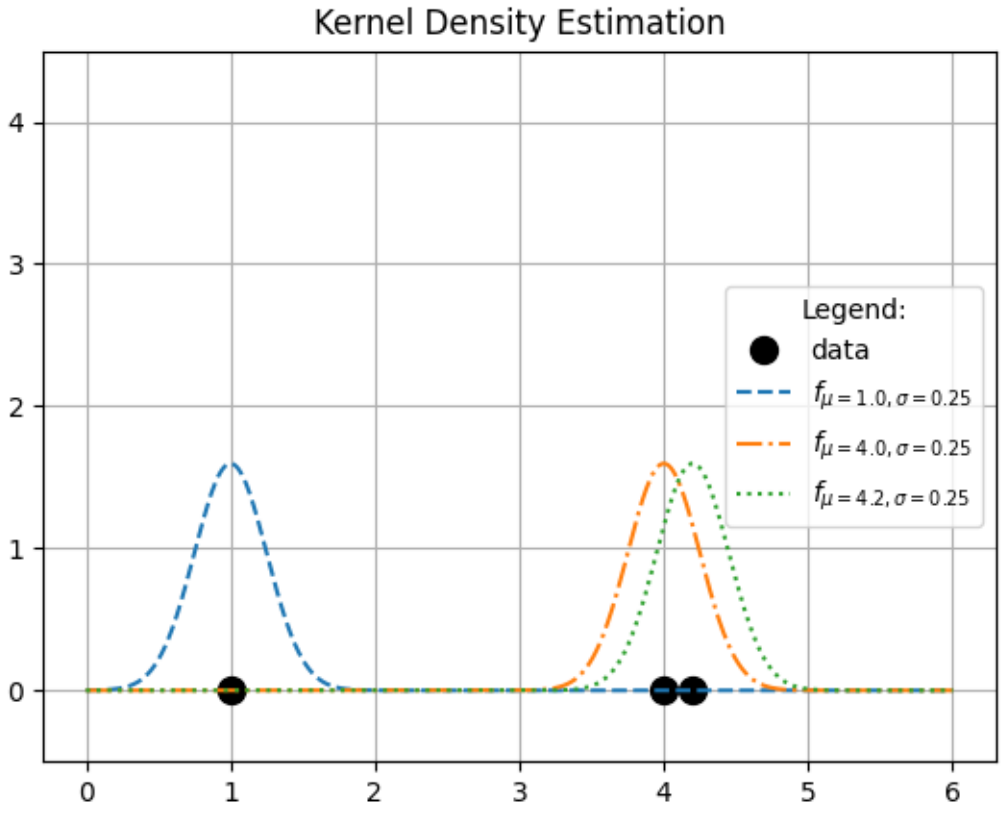
(g) Ridge regression

- ◯ Yes.
- ◯ No.

(h) Permutation feature importance

- ◯ Yes.
- ◯ No.

5. Here is a graph of 1D data $\{\mathbf{x}_i\} = \{x_i\} = \{1, 4, 4.2\}$ and corresponding Gaussian curves $\{f_{\mu=x_i,\sigma=b}(x)\}$ made with bandwidth $b = 0.25$.



(a) Supposing the data were randomly sampled from some population, use kernel density estimation to estimate the population's probability density $f(x)$ at $x = 4.1$.

Based on the plot, the estimate is $\hat{f}_{b=0.25}(4.1) \approx$ ____.

(b) Estimate the density at $x = 2$.

Based on the plot, the estimate is $\hat{f}_{b=0.25}(2) \approx$ ____.

(c) On the figure above, draw the estimated density function over the interval $[0, 6]$.

6. Consider the use of gradient boosting to train a regression model on the following data:

x	y
1	2
2	1
3	3

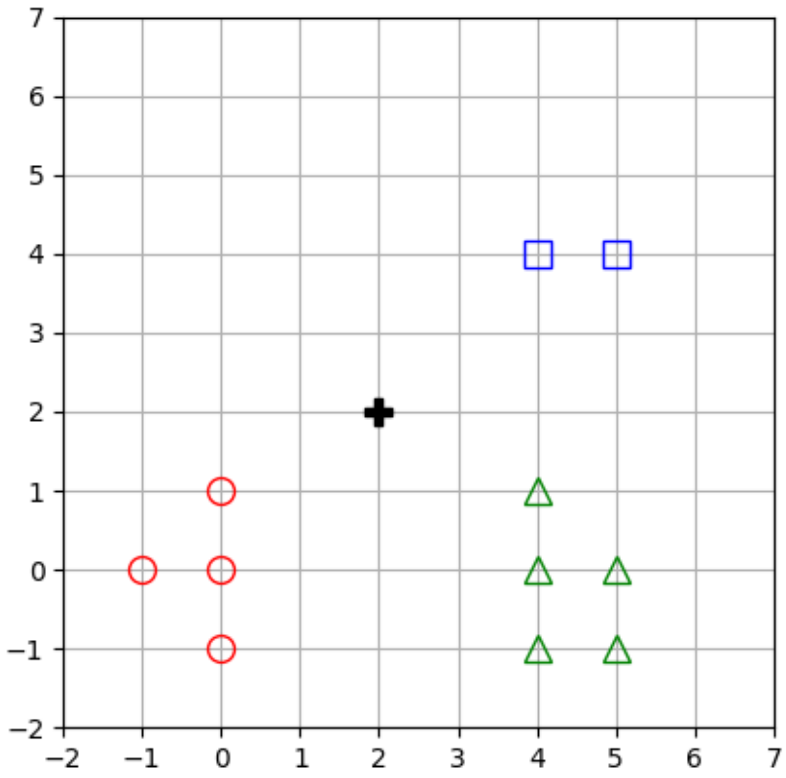
(a) What prediction $\hat{y} = f_0(x)$ is given by the first model at $x = 2$?

$\hat{y} =$ ☐

(b) What data are used to train the second model $\hat{y} = f_1(x)$?

x	y

7. Consider a one-vs.-rest hard-margin linear SVM classifier trained on the following data depicted by circles, squares, and triangles:

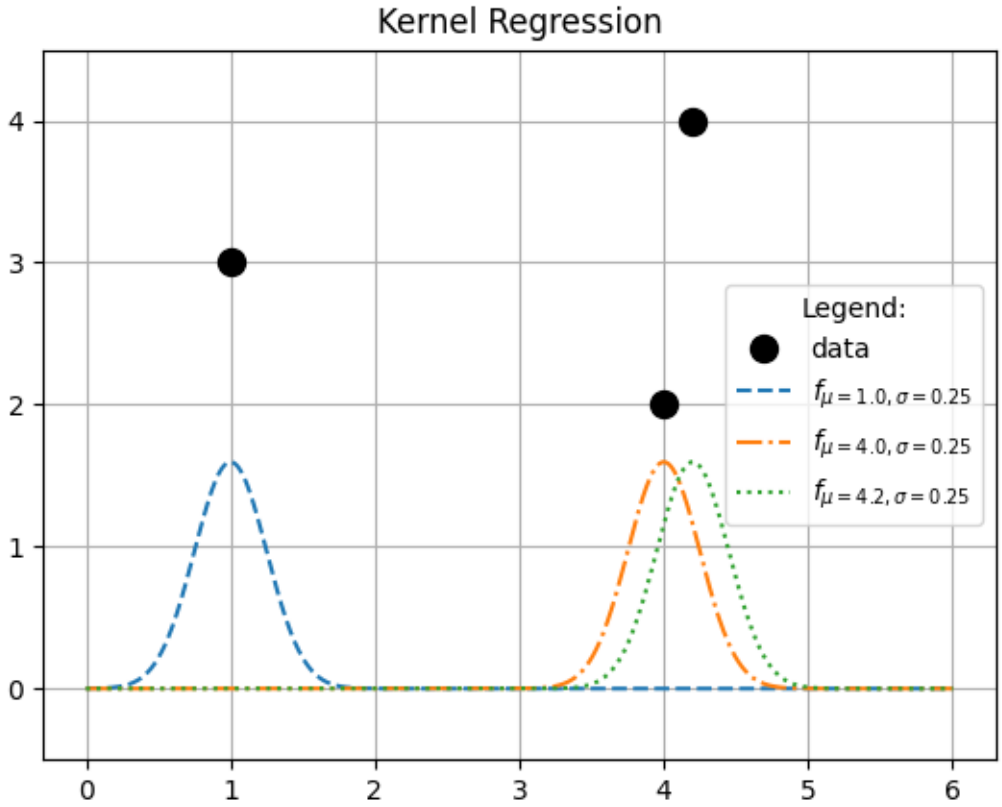


(a) On the graph above, draw the three binary classifiers required by this method.

(b) How does this classifier classify the point indicated by "+"?

- ◯ circle
- ◯ square
- ◯ triangle

8. Here is a graph of the data set $\{(\mathbf{x}_i, y_i)\} = \{(x_i, y_i)\} = \{(1,3),(4,2),(4.2,4)\}$ (here each $\mathbf{x}_i$ is a 1D x_i) along with corresponding Gaussian curves $\{f_{\mu=x_i,\sigma=b}(x)\}$ made with bandwidth $b = 0.25$:



(a) Use kernel regression to estimate $y = f(x)$ for $x = 4.1$.

Based on the plot, the estimate is $\hat{y} \approx$ _____.

(b) Estimate $y = f(x)$ for $x = 1$.

Based on the plot, the estimate is $\hat{y} \approx$ _____.

(c) On the figure above, draw the estimated regression function over the interval $[0, 6]$.

9. Consider the following questions about model assessment.

(a) Consider a classifier trained on examples $(\mathbf{x}, y)$ in the first two columns of the table below that makes the predictions on training data in the third column.

$\mathbf{x}$	y	$\hat{y}$
(0, 2)	1	1
(1, 1)	0	0
(2, 3)	0	0
(4, 1)	0	1

Complete the corresponding confusion matrix:

	predicted $\hat{y}$	
actual y	0	1
0		
1		

(b) The classifier is evaluated on unseen test data yielding this confusion matrix:

	predicted $\hat{y}$	
actual y	0	1
0	3	2
1	1	4

What is the precision on the test data?

(c) What is the recall on the test data?

(d) What is the accuracy on the test data?

(e) Consider a classifier for which we can have either false positive rate 0 with true positive rate 0.5 or false positive rate 0.5 with true positive rate 1. What is the AUC?

(f) For a classifier with TPR = 0 and FPR = 1, what is the AUC?

(g) For each situation, indicate whether precision or recall should be optimized:

i. A bank is doing fraud detection where a fraudulent transaction ("positive") that is missed is expensive but a valid transaction labeled fraudulent is inexpensive.
- ◯ Precision
- ◯ Recall

ii. A doctor is screening patients for a disease in which an ill patient ("positive") infects others and dies if the disease is not diagnosed.
- ◯ Precision
- ◯ Recall

iii. A marketing campaign invests considerable expense in a prospective customer when it classifies that customer as likely to make a purchase ("positive").
- ◯ Precision
- ◯ Recall