STAT 451 Exam2

Summer 2025

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

x	y
1	7
2	6
3	5

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

   $\hat{y} =$ [ ]

   ANSWER:

   $\hat{y} = 6$ because $f_0(x)$ is a constant model given by $f_0(x) = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{3}(7+6+5) = 6$.

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

x	y

   ANSWER:

x	y
1	1
2	0
3	-1

   f_1 is trained on the original data after replacing the y values with the residuals with respect to the first model, that is with $\{e_i = y_i - f_0(x_i)\}$. Since $f_0(x) = 6$ for all x, $\{e_i\} = \{7-6, 6-6, 5-6\} = \{1, 0, -1\}$.

2. Principal Components Analysis (PCA) is a popular method for linear dimensionality reduction. Suppose that we have a dataset with $N = 10$ examples, each composed of $D = 7$ features. We want to find a 2-dimensional subspace via PCA that retains the most feature from this dataset.

   (a) In the first computational step, the *Gram matrix* $X^T X$ is found. Its shape is _______ rows by _______ columns.

   ANSWER: X is $X_{N\times D} = X_{10\times 7}$, so $X^T X$ is $X^T_{7\times 10} X_{10\times 7} = (X^T X)_{7\times 7}$, i.e. 7 rows by 7 columns.

   (b) Mark each statements about PCA as True or False.

i. It seeks a dataset with a smaller number of features.
ANSWER: True

ii. It seeks a dataset with a smaller number of examples.
ANSWER: False

iii. For each pair of principal components $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{a} \cdot \mathbf{b} = 0$.
ANSWER: True

iv. For each pair of principal components $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{a} + \mathbf{b} = 0$.
ANSWER: False

v. Any subset of three principal components explain most of the variability in the original data.
ANSWER: False

vi. PCA is useful for feature selection.
ANSWER: False

vii. Rescaling data is unnecessary before PCA because principal components have unit length whether or not the data are rescaled first.
ANSWER: False

3. Consider using k-means on the unsupervised 1D dataset $\{x\} = \{1, 3, 5, 10, 12\}$ to create $k = 2$ clusters. Suppose the two initial randomly-chosen cluster centroids are $c_1 = 3$ and $c_2 = 5$.

(a) What are the centroids after the first iteration of k-means?
$c_1 =$ _____ and $c_2 =$ _____.
ANSWER: 1 and 3 are closest to $c_1 = 3$, so the new $c_1 = \frac{1}{2}(1 + 3) = 2$.
5, 10, and 12 are closest to $c_2 = 5$, so the new $c_2 = \frac{1}{3}(5 + 10 + 12) = 9$.

(b) What are the centroids after the second iteration?
$c_1 =$ _____ and $c_2 =$ _____.
ANSWER: 1, 3, and 5 are closest to $c_1 = 2$, so the new $c_1 = \frac{1}{3}(1 + 3 + 5) = 3$.
10 and 12 are closest to $c_2 = 9$, so the new $c_2 = \frac{1}{2}(10 + 12) = 11$.

4. For each part, select all that apply.

(a) Which model(s) can be trained using gradient decent?
- ○ kNN
- ○ logistic regression ●
- ○ linear regression ●
- ○ decision tree

(b) Which model(s) require looping through the data at least once for training?
- ○ kNN
- ○ logistic regression ●
- ○ linear regression ●
- ○ decision tree ●
- ○ SVM ●

(c) Which model(s) have a fixed number of parameters?

○ kNN ●

○ logistic regression ●

○ linear regression ●

○ decision tree

○ SVM ●

(d) Which model(s) can exhibit improved performance by scaling features?

○ kNN ●

○ logistic regression ●

○ linear regression ●

○ decision tree

○ SVM ●

(e) Which model(s) requires the least amount of time to train?

○ kNN ●

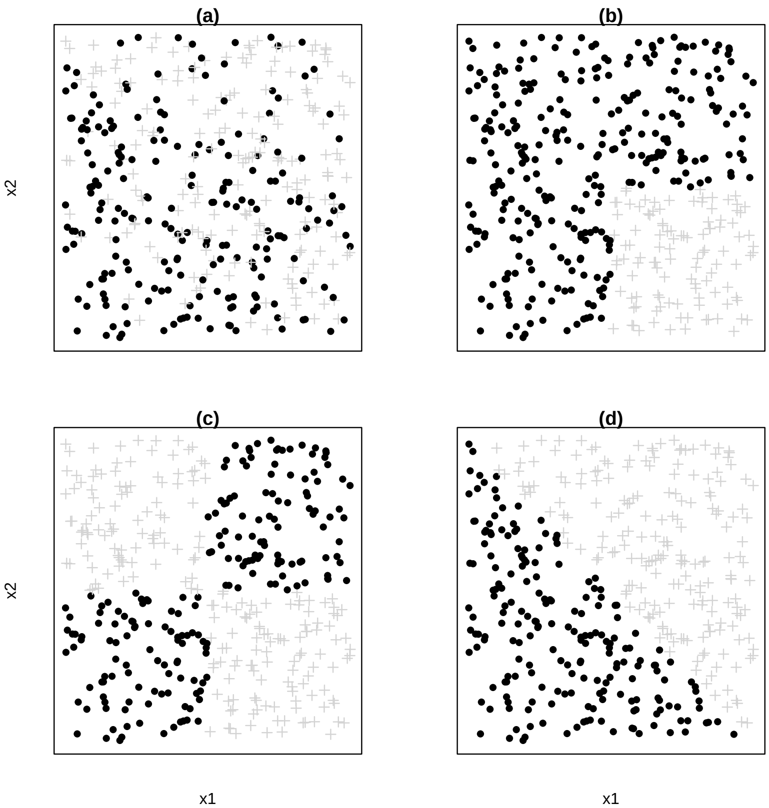
○ logistic regression

○ linear regression

○ decision tree

○ SVM

5. Match the four data sets to the four classification models to produce the best overall accuracy. Use each model once. The features are $\mathbf{x}_1$ and $\mathbf{x}_2$. The points are colored by class: $y = 1$ is a gray "+" and $y = 0$ is a black dot.



(a) ○ kNN
○ logistic regression
○ SVM with your choice of linear or RBF kernel ●; use RBF kernel for nonlinear data
○ decision tree with maximum depth 3

(b) ○ kNN
○ logistic regression
○ SVM with your choice of linear or RBF kernel
○ decision tree with maximum depth 3 ●

(c) ○ kNN ●; note that no decision tree split can reduce entropy and the data are not linearly separable
○ logistic regression
○ SVM with your choice of linear or RBF kernel
○ decision tree with maximum depth 3

(d) ○ kNN
○ logistic regression ●; note that decision tree would require depth much higher than 3
○ SVM with your choice of linear or RBF kernel
○ decision tree with maximum depth 3

6. Consider kernel density estimation on the data $\{x_i\} = \{0, 1, 2\}$.

(a) Decreasing the bandwidth b toward 0 makes the density estimate at $x = 1$ tend toward:

- ◯ 0
- ◯ $\frac{1}{3}$
- ◯ $\frac{1}{2}$
- ◯ 1
- ◯ 3
- ◯ ∞ ANSWER: ●

(b) Decreasing the bandwidth b toward 0 makes the density estimate at $x = 1.5$ tend toward:

- ◯ 0 ANSWER: ●
- ◯ $\frac{1}{3}$
- ◯ $\frac{1}{2}$
- ◯ 1
- ◯ 3
- ◯ ∞

7. Consider the use of bagging applied to classification decision trees of depth 1 (one decision node and two leaf nodes per tree). A training data set, on the left, consists of $\{(\mathbf{x}, y)\} = \{(x, y)\}$ because $\mathbf{x}$ has only one feature, x. It is followed by $B = 3$ bootstrap resamples created by sampling with replacement from the training data.

Training data

x	y
1	1
2	0
3	1
4	0

Resample #1

x	y
1	1
2	0
4	0
4	0

Resample #2

x	y
1	1
1	1
3	1
4	0

Resample #3

x	y
1	1
1	1
2	0
2	0

Consider making a prediction for $\mathbf{x} = 2$.

(a) What prediction is made by the tree trained on Resample #1? $\hat{y} =$ ☐

ANSWER:

The tree uses threshold $t = 1.5$ and predicts $\hat{y} = 0$.

(b) What prediction is made by the tree trained on Resample #2? $\hat{y} =$ ☐

ANSWER:

The tree uses threshold $t = 3.5$ and predicts $\hat{y} = 1$.

(c) What prediction is made by the tree trained on Resample #3? $\hat{y} =$ [ ]

ANSWER:

The tree uses threshold $t = 1.5$ and predicts $\hat{y} = 0$.

(d) What prediction is made by this bagging classifier? $\hat{y} =$ [ ]

ANSWER:

The bagging classifier predicts $\hat{y} = 0$, the most frequent of the $B = 3$ predictions.

8. (a) Mark each statement as True or False.

   i. When false positives are costly, precision is a good assessment metric. True

   ii. When false negatives are costly, recall is a good assessment metric. True

   iii. A police officer running a blood alcohol test on a driver wants high precision to avoid missing a drunk driver. False

   iv. The same police officer wants high recall to avoid arresting a sober driver. False

   (b) Suppose that we want to conduct a 10-fold cross-validation.

   i. How many validation sets will we have in total? _______
   ANSWER: 10, one for each fold

   ii. How many distinct training sets will we have in total? _______
   ANSWER: 10.

   (c) Suppose we conduct ordinary least-squares linear regression on a data set and find it works well on training data but poorly on test data. What methods could we expect to address the problem? Select all that apply.

   i. Use lasso regression on the data. Yes

   ii. Use ridge regression on the data. Yes

   iii. Use logistic regression on the data. No

   iv. Run PCA on the data. Then train the linear model on PCA-reduced-dimension data. Yes

   v. Cluster the examples. Then train the linear model on cluster centers. No

   (d) For each situation, indicate which hyperparameter search strategy, $G =$ grid search or $R =$ random search, is more likely to be successful. Suppose computation time is limited.

   i. A model has two hyperparameters. The first takes one of two string values and the other takes one of three numeric values.
   ANSWER: G = grid search.

   ii. A model has two hyperparameters. The first takes a floating-point number in the interval $[0, 1]$ while the second takes an integer in the range $[0, 1000]$.
   ANSWER: R = random search

9. Mark each statement True or False.

(a) In kernel density estimation we use a weighted average of the y's, where the weights come from Gaussians centered at the x's, to determine the height of an approximate density curve at each x.

◯ True

◯ False ANSWER: ●; there are no y's in KDE

(b) One-hot encoding of a categorical feature with five categories produces $2^5 = 32$ new binary features (or 31 if we use `drop_first = True`).

◯ True

◯ False ANSWER: ●

(c) When principal component analysis retains some number p of principal components of a D-dimensional data set (where $0 < p < D$), those p components are the subset of p features that explain the most variance in the original data.

◯ True

◯ False ANSWER: ●; while each principal component is a linear combination of features, it is not typically a feature.

10. Suppose we use a stacking model to do multiclass classification such that:

- The data are:

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

- The first base model is (multinomial) logistic regression.
- The second base model is k-NN.
- The meta-model is a decision tree.

(a) The first model is trained on how many features?

◯ 0

◯ 1

◯ 2 ANSWER: ● (It is trained on the $\mathbf{x}$'s.)

◯ 3

◯ 4

◯ 5

◯ 6

(b) The second model is trained on how many features?

◯ 0

◯ 1

- ◯ 2 ANSWER: ● (It is trained on the $\mathbf{x}$'s.)
- ◯ 3
- ◯ 4
- ◯ 5
- ◯ 6

(c) The meta-model is trained on how many features?

- ◯ 0
- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ 5
- ◯ 6 ANSWER: ● (Its training examples are concatenated output probability vectors from the two base models, each of which has length 3, for a total length of 6.)