DSO530 Statistical Learning Methods

Lecture 4: Classification III

Dr. Xin Tong
Department of Data Sciences and Operations
Marshall School of Business
University of Southern California
xint@marshall.usc.edu

Linear discriminant analysis

- Linear discriminant analysis (LDA) is a model that appears often in the statistics/ML literature for its nice mathematical properties to analyze
- We skip its details here because it usually has a similar empirical performance to logistic regression, and in practice, it has far less popular compared to logistic regression
- However, it is worth to know that from the modeling perspective, LDA and logistic regression are two different approaches.
- What is an LDA model:

$$X|(Y=0) \sim \mathcal{N}(\mu_0, \Sigma) \quad \text{vs.} \quad X|(Y=1) \sim \mathcal{N}(\mu_1, \Sigma)$$

- Key: two class normal; different means, but same covariance matrix
- How about model fitting? The means and covariance are unknown parameters. We use the so-called **plug-in approach** (not required).
- LDA gives linear decision boundary.
- Recall the logistic regression model. Can you name one difference between the LDA model and the logistic regression model?

Bayes' Theorem

proof: $P(A|B) = P(A \text{ and } B)/P(B)$ ① $P(B|A) = P(A \text{ and } B)/P(A)$ ② ② $\Rightarrow P(A \text{ and } B) = P(B|A) \cdot P(A)$ ← ③

③ & ① gives Bayes' Theorem. ☺

$P(B) = P(B \text{ and } A) + P(B \text{ and } A^c) = P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)$

- Let A and B be two events. Then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- An example involving Bayes' Theorem. Suppose in the population, there are 5% of the people using drug D. Scientists have developed a test to identify the users of this drug. This test has false negative rate 10% and false positive rate 20%. Then given that a person is tested positive, what is the probabilty that he/she is a drug D user?

$P(\text{user}) = 5\%$, $P(\text{user}^c) = 95\%$. $P(\text{pos}) = ?$

$P(\text{neg}|\text{user}) = 10\%$, $P(\text{pos}|\text{user}) = 90\%$

$P(\text{pos}|\text{non user}) = 20\%$

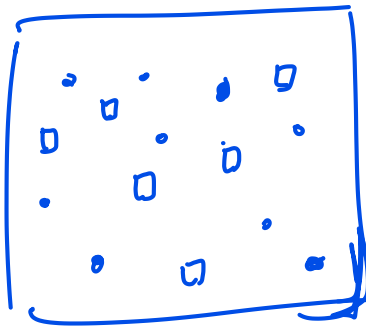
Q: $P(\text{user}|\text{pos}) \overset{?}{=} 19\%$

Bayes' Theorem (another version, optional)

- Suppose $Y \in \{0, 1\}$, and $X|(Y = 0)$ and $X|(Y = 1)$ are continous
- $P(Y = 1|X = x) = \frac{f_{X|Y=1}(x) \cdot P(Y=1)}{f_X(x)} = \frac{f_{X|Y=1}(x) \cdot P(Y=1)}{f_{X|Y=1}(x) \cdot P(Y=1) + f_{X|Y=0}(x) \cdot P(Y=0)}$
- The above allows one to compute $P(Y = 1|X = x)$ from the LDA assumption

K-nearest neighbors (KNN)

supervised learning method
not k-means.



Goal: classify pt to either □ or ●.
① Calculate dist b/w pt and every pt in training set.
② Find k pts in training set who are closest to new pt. (default: k)
③ Find most frequent class among those k pts
④ Assign pt to most frequent class.

- 'KNN': to predict class label for an observation $X = x$, the K training observations that are closest to x are identified. Then x is assigned to the class to which the plurality of these observations belong
- Why did we say "plurality" instead of *majority*?
- Can predict the class label for any $x \in R^p$ (including the training observations) in this way
- Special cases: $K = 1$ and $K = n$ (n is the training sample size)

training error:

$K = 1$: 0% as the closest pt to a pt is itself

$K = n$: 6/15 according to 6□ & 9● above

K-nearest neighbors (KNN)

① KNN does not assume probabilistic model.
② Has non linear decision boundary.

- KNN is a very different approach to classification. Can you name some differences from logistic regression and LDA?
- On training dataset, which choice of K gives the smallest error? Will you choose this K? No, due to overfitting. $k=1$
- In general, as K increases, how will training error change? What about test error?
- Does KNN have linear decision boundaries? No

Test error is not good for any situation.

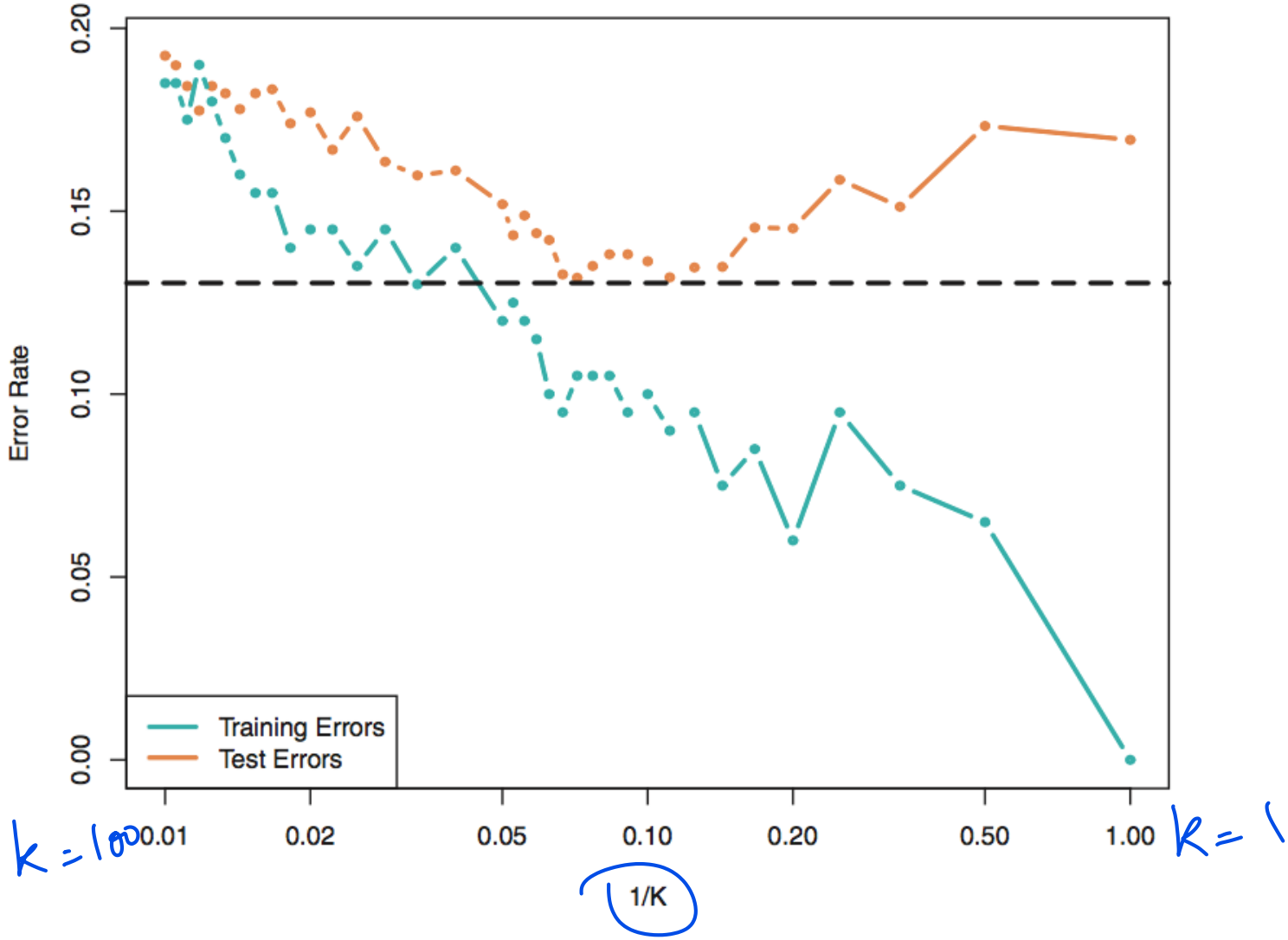


FIGURE 2.17. *The KNN training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) on the data from Figure 2.13, as the level of flexibility (assessed using $1/K$) increases, or equivalently as the number of neighbors K decreases. The black dashed line indicates the Bayes error rate. The jumpiness of the curves is due to the small size of the training data set.*

```
import numpy as np
import pandas as pd
url="https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
# Assign colum names to the dataset
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'Class']
# Read dataset to pandas dataframe
dataset = pd.read_csv(url, names=names); dataset.head()
```

```
##    sepal-length  sepal-width  petal-length  petal-width        Class
## 0           5.1          3.5           1.4          0.2  Iris-setosa
## 1           4.9          3.0           1.4          0.2  Iris-setosa
## 2           4.7          3.2           1.3          0.2  Iris-setosa
## 3           4.6          3.1           1.5          0.2  Iris-setosa
## 4           5.0          3.6           1.4          0.2  Iris-setosa
```

- This is perhaps the best known database to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

```
np.unique(dataset["Class"])
```

```
## array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

```
dataset.info()
```

```
## <class 'pandas.core.frame.DataFrame'>
## RangeIndex: 150 entries, 0 to 149
## Data columns (total 5 columns):
##  #   Column        Non-Null Count  Dtype
## ---  ------        --------------  -----
##  0   sepal-length  150 non-null    float64
##  1   sepal-width   150 non-null    float64
##  2   petal-length  150 non-null    float64
##  3   petal-width   150 non-null    float64
##  4   Class         150 non-null    object
## dtypes: float64(4), object(1)
## memory usage: 6.0+ KB
```

```
dataset.describe()
```

```
##        sepal-length  sepal-width  petal-length  petal-width
## count    150.000000   150.000000    150.000000   150.000000
## mean       5.843333     3.054000      3.758667     1.198667
## std        0.828066     0.433594      1.764420     0.763161
## min        4.300000     2.000000      1.000000     0.100000
## 25%        5.100000     2.800000      1.600000     0.300000
## 50%        5.800000     3.000000      4.350000     1.300000
## 75%        6.400000     3.300000      5.100000     1.800000
## max        7.900000     4.400000      6.900000     2.500000
```

- The four predictors are on the same scale
- In the slides, we will not use feature rescaling. After you go home, please try the feature rescaling step and see if the result becomes better.

```
from sklearn.model_selection import train_test_split
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, 4].values
X_train, X_test, y_train, y_test=\
  train_test_split(X, y, test_size=0.20, random_state=5, stratify=y)

from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors=5)
classifier.fit(X_train, y_train)
```

```
## KNeighborsClassifier()
```

- There are many default parameters here. We just pay attention to `weights`.

```
y_pred = classifier.predict(X_test)
np.mean(y_pred != y_test)
```

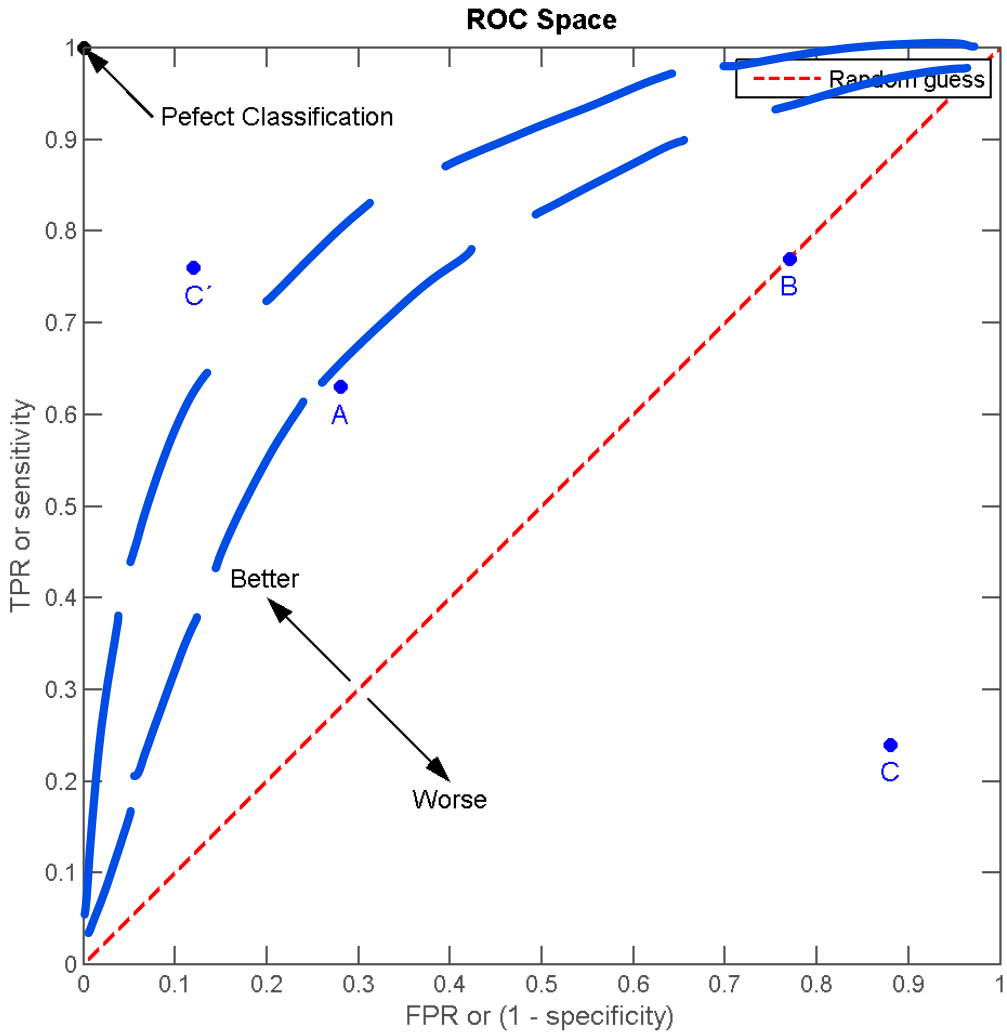
```
## 0.03333333333333333
```

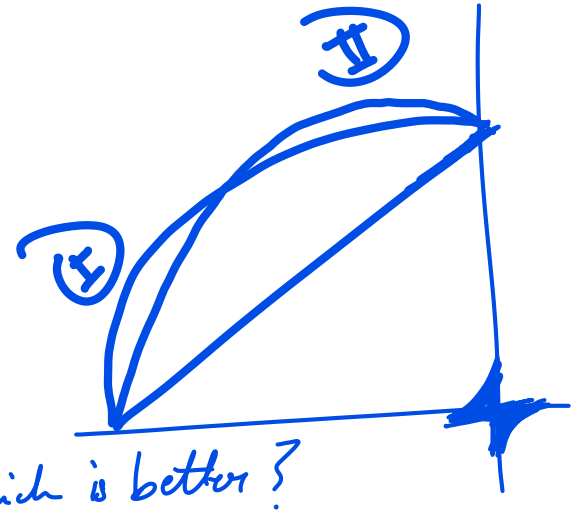
- After you go home, try practice different settings, such as `random_state=100`, `n_neighbors = 3`, `n_neighbors = 7`

Understanding different errors

- First recall type I/II error from the confusion matrix
- specificity = 1 - type I error $1 - \frac{b}{a+b} = a/a+b$
- power (a.k.a. sensitivity; a.k.a. recall) = 1 - type II error $1 - \frac{c}{c+d} = \frac{d}{c+d}$
- precision = d/(b+d)
- Sometimes, type I error is calledd false positive rate (fpr); 1- type II error is called true positive rate (tpr)

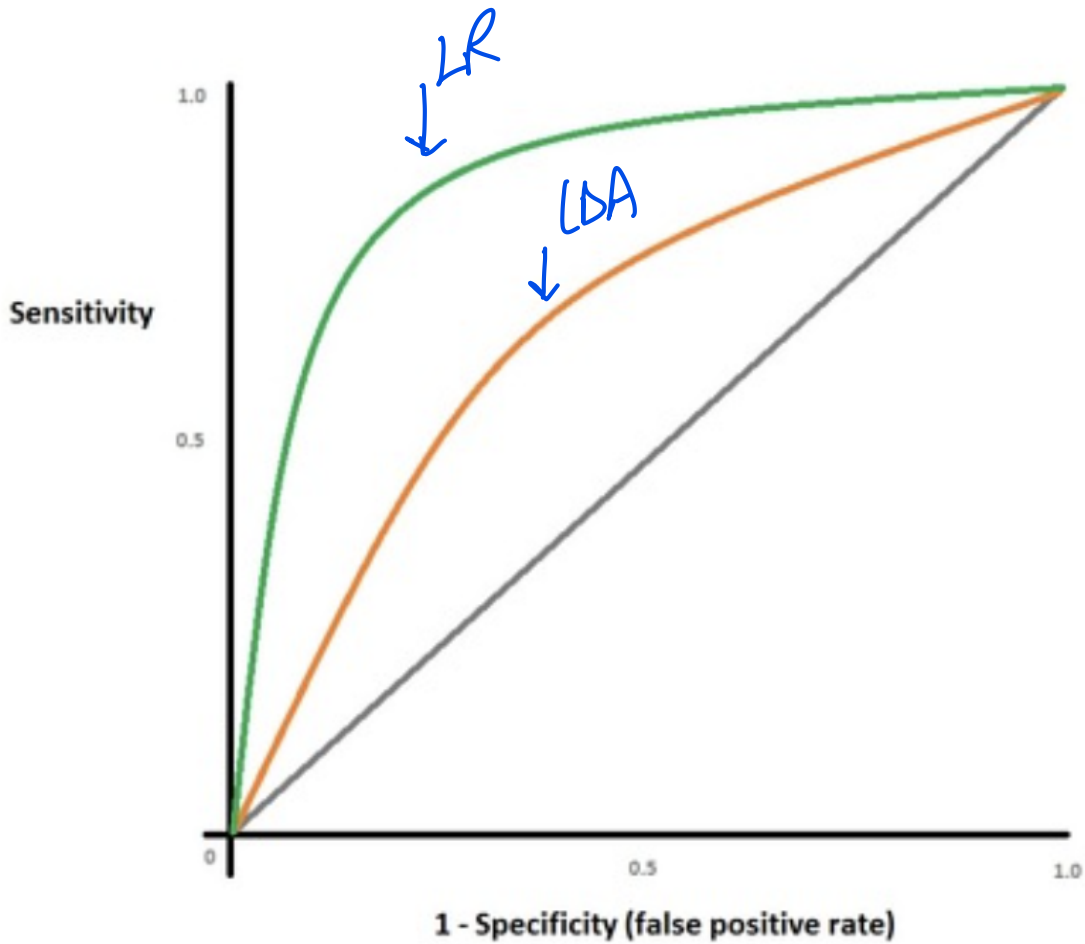
Receiver operating characteristic (ROC) space





which is better?
the curve with more area under the curve.
Also determine priority of type ① & ② errors!

Receiver operating characteristic (ROC) curves



- Which curve is better? One way to judge: area under the curve (AUC)

Data science vs. statistics

- Optional and for fun only
- https://www.youtube.com/watch?v=uHGICi9jOWY&feature=youtu.be&fbclid=IwAR1tTAIEhgUrHk815BIMj0PxC8XYRLz62pA_V6C___xSWXk3ZAF8uoj9I0JQ
- In some sense, the first half of DSO 530 is more of statistics, and the second half is more data science (machine learning)

Review question 1

Recall that a linear classifier means that its decision boundary is linear. Please show that logistic regression gives us linear classifiers. Concretely, show the classifier

$$1\left(\frac{e^{3+9x_1+5x_2}}{1+e^{3+9x_1+5x_2}} > 0.8\right)$$

has a linear decision boundary.

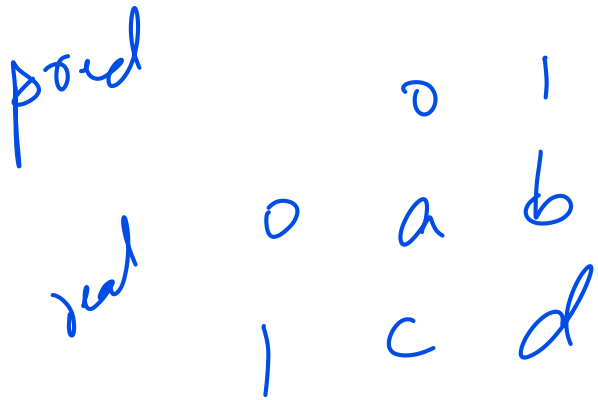
$3+9x_1+5x_2 = \log(4)$

Review question 2

$$X \mid Y=0 \sim N(\mu_0, \Sigma)$$
$$X \mid Y=1 \sim N(\mu_1, \Sigma)$$

What is the linear discriminant analysis model?

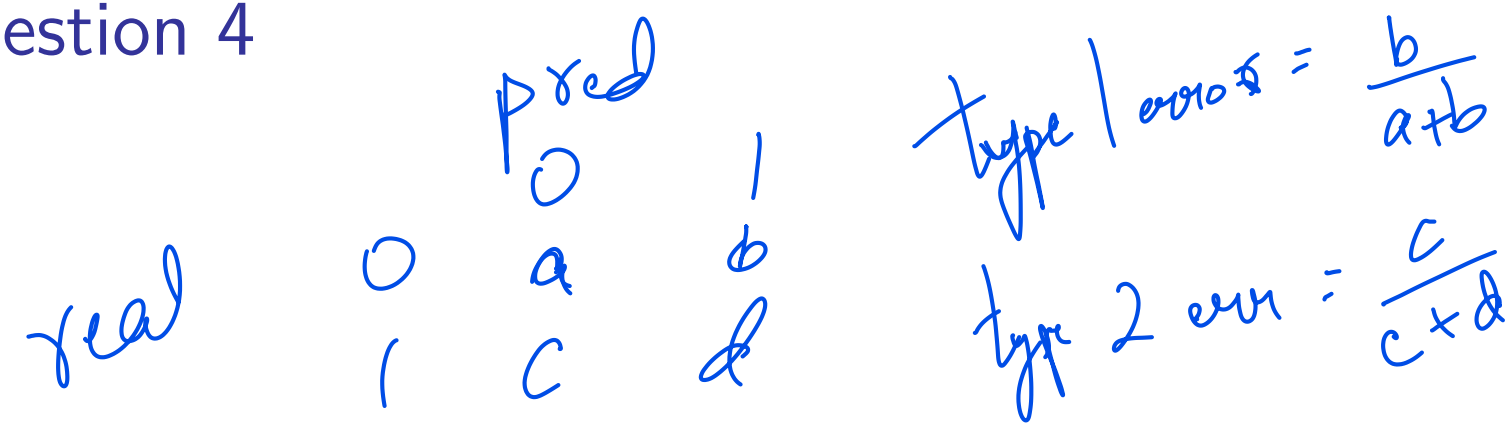
Review question 3



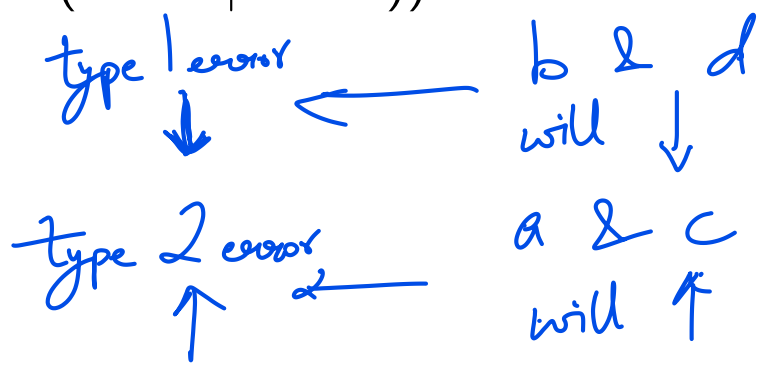
Is the following claim TRUE or FALSE: "As I collect more and more observations, I can usually achieve (close to) perfect classification."

False

Review question 4



Is the following claim TRUE or FALSE: "For logistic regression, when I increase the decision threshold (on $P(Y = 1|X = x)$) from 0.5 to 0.6, the type I error will increase as well"



Review question 5

receiver operating characteristic.

What does the "ROC" in ROC curve stand for?

① AUC
② Compare type 1 error for the same type 2 error

Review question 6

How to use ROC curves to compare classification models?

Review question 7

If ROC curve for model A intersects with ROC curve for model B, what can you say about these two models?

for type I error range, I model dominates for another range other model dominates