COSC 3337
Data Science I
Section 14623

Exam I Review

Instructor: Jingchao Ni
Fall 2024

About Exam I

- Time and location: Oct. 14 (Monday) 2:30PM, SEC 104
- Length: 90 Minutes
- Exam Materials:
  - You can use the lectures notes/slides
  - You can also reference reading material distributed as part of your reading assignment or suggested textbooks for the course
  - Use calculator or python for some computations
  - NO GENERAL AI RESOURCES SHOULD BE USED
- Exam Contents: 11 questions (100 points)
  - Data Preprocessing
  - Exploratory Data Analysis
  - Classification

Question #1

Data Preprocessing

- Given a document

> Random walk is an algorithm for web clustering and hyperlink mining. It is an important algorithm in data mining and web mining.

Please form your dictionary of words as attributes and transform the document to a vector with numerical values. (do not include stop words such as “it”, “is”, “an”, “for”, “in”, “and”)

{“random”, “walk”, “algorithm”, “web”, “clustering”, “hyperlink”, “data”, “mining”}

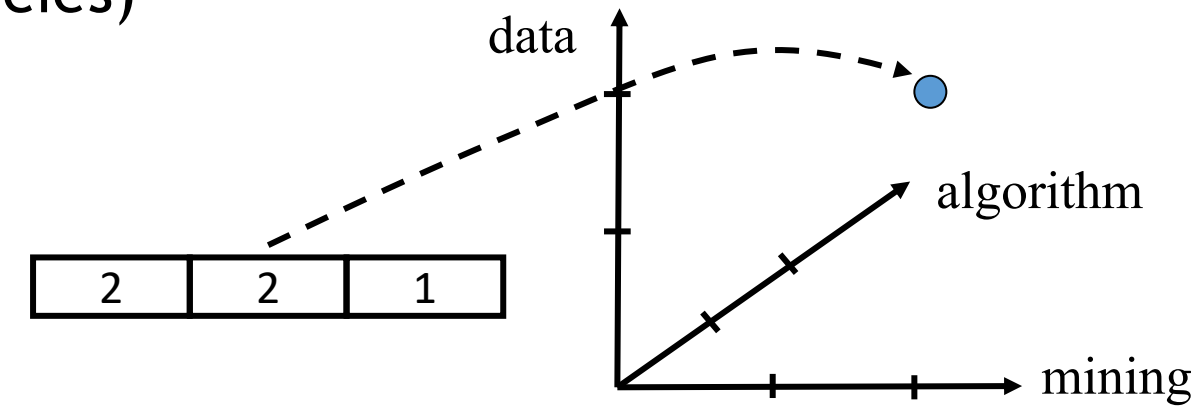
1	1	2	2	1	1	1	3

The numbers represent the frequencies of the respective words.

Mapping into Vector Space

- Many algorithms assume that the input is a vector:
  - Linear regression, SVM, neural networks, etc.
- Text documents not represented as records / tuples
  - "Clustering is one of the generic data mining tasks. One of the most important data mining algorithms"
  - Step 1:Choose attributes (relevant terms / dimensions of vector space)
    - "Data", "mining", and "algorithms"
  - Calculate attribute values (frequencies)
    - Data: 2, mining:2 algorithms: 1
  - Map object to vector in this space

Clustering is one of the generic data mining tasks. One of the most important data mining algorithms . . .



Question #2

Data Preprocessing

- Given a training dataset with 5 records and two categorical attributes

Please transform it to numerical attributes using one-hot representation.

	A	B
1	small	single
2	medium	single
3	large	divorced
4	medium	married
5	large	single

	small	medium	large	single	divorced	married
1	1	0	0	1	0	0
2	0	1	0	1	0	0
3	0	0	1	0	1	0
4	0	1	0	0	0	1
5	0	0	1	1	0	0

Question #3

Data Preprocessing

- Given 10 2-D data points, we want to reduce its dimension to 1-D. Suppose we use PCA, and the learnt principal components are as following. Please compute the 1-D data after dimension reduction. Please include your computation process.

$$\begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 & X_{10} \end{matrix}$$
$$\begin{bmatrix} 1.0 & 2.0 & 2.0 & 4.0 & 3.0 & 2.0 & 1.0 & 5.0 & 3.0 & 5.0 \\ 0.5 & 1.0 & 0.2 & 0.4 & 0.3 & 0.1 & 0.5 & 1.0 & 0.2 & 0.5 \end{bmatrix}$$

$$\begin{matrix} P_1 & P_2 \end{matrix}$$
$$\begin{bmatrix} 2 & 0.1 \\ 10 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 10 \end{bmatrix} \times \begin{bmatrix} 1.0 & 2.0 & 2.0 & 4.0 & 3.0 & 2.0 & 1.0 & 5.0 & 3.0 & 5.0 \\ 0.5 & 1.0 & 0.2 & 0.4 & 0.3 & 0.1 & 0.5 & 1.0 & 0.2 & 0.5 \end{bmatrix}$$

$$= \begin{bmatrix} 7 & 14 & 6 & 12 & 9 & 5 & 7 & 20 & 8 & 15 \end{bmatrix}$$

In Principal Component Analysis (PCA), we choose the principal component that explains the most variance in the data. In this case, $P_1 = \begin{bmatrix} 2 \\ 10 \end{bmatrix}$ is the first principal component, and it explains more variance than $P_2 = \begin{bmatrix} 0.1 \\ 1 \end{bmatrix}$, which is the second principal component.

The reason we choose P_1 instead of P_2 is because PCA orders the components by the amount of variance they explain. P_1, with larger values, likely corresponds to the direction in the data that has the highest variance, making it the most significant component for dimensionality reduction. When reducing to one dimension, we want to retain as much information (variance) as possible, so we project the data onto P_1.

If we had chosen P_2, we would lose more information because P_2 corresponds to the direction with lower variance. Thus, projecting the data onto P_2 would retain less of the original data's structure and information.

Data Preprocessing

- Given 10 2-D data points, we want to reduce its dimension to 1-D. Suppose we use PCA, and the learnt principal components are as following. Please compute the 1-D data after dimension reduction. Please include your computation process.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1.0	2.0	2.0	4.0	3.0	2.0	1.0	5.0	3.0	5.0
0.5	1.0	0.2	0.4	0.3	0.1	0.5	1.0	0.2	0.5

P_1	P_2
2	0.1
10	1

Chose P1 because it has more "variability" (leads to more variability).

$$[2 \quad 10] \times \begin{bmatrix} 1.0 & 2.0 & 2.0 & 4.0 & 3.0 & 2.0 & 1.0 & 5.0 & 3.0 & 5.0 \\ 0.5 & 1.0 & 0.2 & 0.4 & 0.3 & 0.1 & 0.5 & 1.0 & 0.2 & 0.5 \end{bmatrix}$$

$$= [7 \quad 14 \quad 6 \quad 12 \quad 9 \quad 5 \quad 7 \quad 20 \quad 8 \quad 15]$$

$\begin{bmatrix} 2 \\ 10 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.5 \end{bmatrix} \rightarrow 2(1.0) + 10(0.5) = 7$

$\begin{bmatrix} 2 \\ 10 \end{bmatrix} \times \begin{bmatrix} 2.0 \\ 1.0 \end{bmatrix} \rightarrow 2(2.0) + 10(1.0) = 14$

$\begin{bmatrix} 2 \\ 10 \end{bmatrix} \times \begin{bmatrix} 2.0 \\ 0.2 \end{bmatrix} \rightarrow 2(2.0) + 10(0.2) = 6$

$\begin{bmatrix} 2 \\ 10 \end{bmatrix} \times \begin{bmatrix} 4.0 \\ 0.4 \end{bmatrix} \rightarrow 2(4.0) + 10(0.4) = 12$

$\begin{bmatrix} 2 \\ 10 \end{bmatrix} \times \begin{bmatrix} 3.0 \\ 0.3 \end{bmatrix} \rightarrow 2(3.0) + 10(0.3) = 9$

$\begin{bmatrix} 2 \\ 10 \end{bmatrix} \times \begin{bmatrix} 2.0 \\ 0.1 \end{bmatrix} \rightarrow 2(2.0) + 10(0.1) = 5$

$\begin{bmatrix} 2 \\ 10 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 0.5 \end{bmatrix} \rightarrow 2(1.0) + 10(0.5) = 7$

$\begin{bmatrix} 2 \\ 10 \end{bmatrix} \times \begin{bmatrix} 5.0 \\ 1.0 \end{bmatrix} \rightarrow 2(5.0) + 10(1.0) = 20$

$\begin{bmatrix} 2 \\ 10 \end{bmatrix} \times \begin{bmatrix} 3.0 \\ 0.2 \end{bmatrix} \rightarrow 2(3.0) + 10(0.2) = 8$

$\begin{bmatrix} 2 \\ 10 \end{bmatrix} \times \begin{bmatrix} 5.0 \\ 0.5 \end{bmatrix} \rightarrow 2(5.0) + 10(0.5) = 15$

Data Reduction – Dimensionality Reduction

- PCA Example
  - Suppose we have 10 2-D data points
  - We want to find a basis vector to project the 2-D data points to 1-D data points while maintaining their largest variation

$$[p_1 \quad p_2] \times \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} & x_{1,5} & x_{1,6} & x_{1,7} & x_{1,8} & x_{1,9} & x_{1,10} \\ x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} & x_{2,5} & x_{2,6} & x_{2,7} & x_{2,8} & x_{2,9} & x_{2,10} \end{bmatrix}$$

1-D basis vector P (Principal components)

Raw data X

$$= \quad [y_1 \quad y_2 \quad y_3 \quad y_4 \quad y_5 \quad y_6 \quad y_7 \quad y_8 \quad y_9 \quad y_{10}]$$

New data Y

Data Reduction – Dimensionality Reduction

- PCA Example
  - Suppose we have 10 2-D data points
  - We want to find a basis vector to project the 2-D data points to 1-D data points while maintaining their largest variation

$$P \times X = Y$$

K by M matrix M by N matrix K by N matrix

Data Reduction – Dimensionality Reduction

- PCA Computation

```
>>> from sklearn.decomposition import PCA
>>> pca = PCA(n_components=2)
>>> Y = pca.fit_transform(X)
```

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

***class* sklearn.decomposition.PCA(*n_components*=None, **, copy*=True, *whiten*=False, *svd_solver*='auto', *tol*=0.0, *iterated_power*='auto', *n_oversamples*=10, *power_iteration_normalizer*='auto', *random_state*=None)**

components_: ndarray of shape (n_components, n_features)
- Principal axes in feature space

explained_variance_: ndarray of shape (n_components,)
- The amount of variance explained by each of the selected components.

singular_values_: ndarray of shape (n_components,)
- The singular values corresponding to each of the selected components.

Question #4

Exploratory Data Analysis

- Given the values of an attribute of 6 records
  [1, 5, -1, 2, 3, 5]

Please calculate its mean, median, variance, absolute average deviation (AAD), median absolute deviation (MAD), and tell which of AAD, MAD, and variance is the least sensitive to outliers and why.

Mean: $(1 + 5 - 1 + 2 + 3 + 5) / 6 = 2.5$
Median: $(2 + 3) / 2 = 2.5$
Variance: $(1.5^2 + 2.5^2 + 3.5^2 + 0.5^2 + 0.5^2 + 2.5^2) / 6 = 4.58$
AAD: $(1.5 + 2.5 + 3.5 + 0.5 + 0.5 + 2.5) / 6 = 1.83$
MAD: $(1.5 + 2.5) / 2 = 2$

MAD is least sensitive to outliers because it only does not use square to enlarge the difference and only uses median (outliers are usually minimum or maximum).

Stats Calculations

Sunday, October 13, 2024 7:39 PM

Exploratory Data Analysis

- Given the values of an attribute of 6 records
  [1, 5, -1, 2, 3, 5]

Please calculate its mean, median, variance, absolute average deviation (AAD), median absolute deviation (MAD), and tell which of AAD, MAD, and variance is the least sensitive to outliers and why.

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

$$\text{variance}(x) = s_x^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \bar{x})^2$$

standard_deviation(x)= s_x

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m} |x_i - \overline{x}|$$

$$\text{MAD}(x) = median\Big(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\Big)$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

Mean: $(1+5-1+2+3+5)/6 = 2.5$

Median: $[-1,1,2,3,5,5]$

$(2+3)/2 = 2.5$

Variance: $(x_i - \bar{x})^2$ $\bar{x}=2.5$ $m=6$

$1-2.5 = -1.5$
$5-2.5 = 2.5$
$-1-2.5 = -3.5$
$2-2.5 = -0.5$
$3-2.5 = 0.5$
$5-2.5 = 2.5$

→ $((-1.5)^2+(2.5)^2+(-3.5)^2+(-0.5)^2+(0.5)^2+(2.5)^2)/6 = 4.58$

AAD: $|x_i - \bar{x}|$ $\bar{x}=2.5$ $m=6$

$|1-2.5| = 1.5$
$|5-2.5| = 2.5$
$|-1-2.5| = 3.5$
$|2-2.5| = 0.5$
$|3-2.5| = 0.5$
$|5-2.5| = 2.5$

→ $(1.5+2.5+3.5+0.5+0.5+2.5)/6 = 1.83$

MAD: $|x_i - \bar{x}|$ $\bar{x}=2.5$ ← Median

Median = 2.5

$|1-2.5| = 1.5$
$|5-2.5| = 2.5$
$|-1-2.5| = 3.5$
$|2-2.5| = 0.5$
$|3-2.5| = 0.5$
$|5-2.5| = 2.5$

→ order in increasing order → $[0.5, 0.5, 1.5, 2.5, 2.5, 3.5]$

↑ Median: $(1.5+2.5)/2 = 2$

REMEMBER: Median is least sensitive to outliers.

Measures of Location: Mean and Median

- The mean is the most common measure of the location of a set of points
- However, the mean is very sensitive to outliers
- Thus, the median or a trimmed mean is also commonly used

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Measures of Spread: Range and Variance

- Range is the difference between the max and min
- The variance or standard deviation

0, 2, 3, 7, 8 → $\bar{x} = 4$

$$\text{variance}(x) = s_x^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2$$

9.2

standard_deviation(x)= s_x

3.3

- However, this is also sensitive to outliers, so that other measures are often used

2.8

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

(Mean Absolute Deviation) [Han]
(Absolute Average Deviation) [Tan]

3

$$\text{MAD}(x) = median\left(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\right)$$

(Median Absolute Deviation)

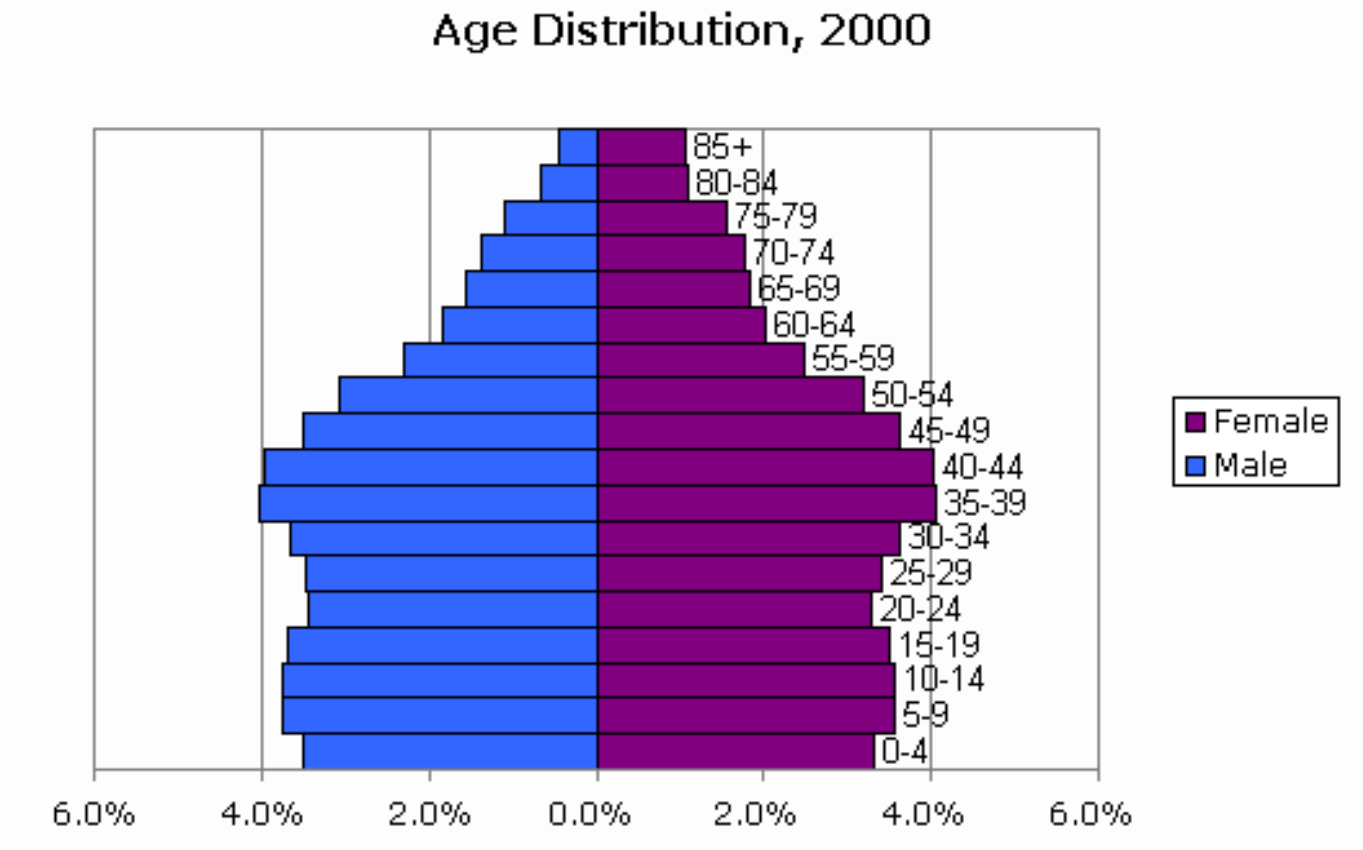
5

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

Question #5

Exploratory Data Analysis

- Interpret the following 2 histograms and their relationships which describe the male and female age distribution in the US, based on Census Data.



Exploratory Data Analysis

- Both histograms:
  - curves are continuous with no gaps or outliers, and somewhat smooth,
  - bimodal with 2 modalities not well separated maxima at 5-19 and 35-44,
  - values significantly drop beyond age 55 → skewed distribution

- Comparison:
  - curves are somewhat similar until age 55 (although there are more males initially); e.g. shape of the density function and the 2 local maxima match
  - however, the decline in the male curve is significantly steeper: women live longer

Interpretation of Histogram

- What is the type of the attribute? Positive real numbers
- What is the mean value?
- Is there a lot of spread or not (compute the standard deviation)? Not much
- Is the distribution unimodal (one hill or no hill)) or multi-modal (multiple hills)? One hill or two hills, depending on how you interpret the data. The second hill is not very well separated; therefore I would say unimodal.
- Is the distribution skewed (e.g. compare mean with median)? Only very mildly skewed
- Are there any outliers? Yes values above 45 …
- Are there any duplicate values?
- Are there any gaps in the attribute value distribution? Yes two gaps: 1)… 2)…
- Characterize the shape of the density function! Bell Curve

Answers are NOT for the histogram in the previous slide.
FOCUS on the questions.

Question #6

Classification

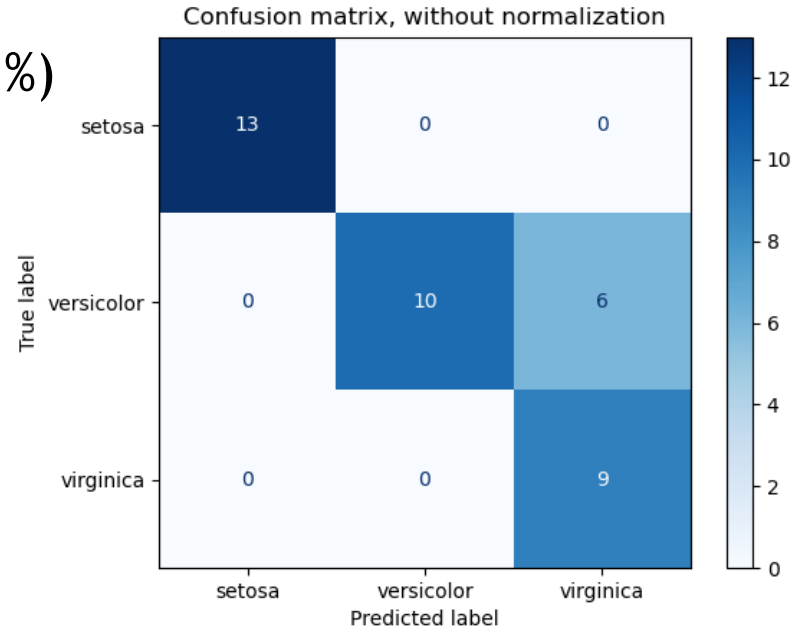
- The following confusion Matrix of a classification model of the IRIS flower dataset is given below.
  What is the accuracy and classification error of the classification model?
  What is its precision for class virginica? What is its recall for class virginica?

Accuracy: (13 + 10 + 9) / (13 + 10 + 9 + 6) = 0.84 (or 84.21%)

Classification error: 1 – 0.84 = 0.16 (or 15.79%)

Precision: 9 / 15 = 0.6 (or 60%)

Recall: 9 / 9 = 1 (or 100%)



	Predicted as Class 1	Predicted as Class 2	Predicted as Class 3
Actual Class 1	**a** (TP for Class 1)	b	c
Actual Class 2	d	**e** (TP for Class 2)	f
Actual Class 3	g	h	**i** (TP for Class 3)

1. **Precision for Class 1**: Precision measures how many of the instances predicted as Class 1 are actually Class 1. The formula is:

$$\text{Precision}_{Class1} = \frac{a}{a + d + g}$$

   This is the ratio of true positives for Class 1 to all instances predicted as Class 1.

2. **Recall for Class 1**: Recall measures how many of the actual Class 1 instances were correctly predicted as Class 1. The formula is:

$$\text{Recall}_{Class1} = \frac{a}{a + b + c}$$

   This is the ratio of true positives for Class 1 to all actual Class 1 instances.

3. **Overall Accuracy**: Accuracy is the proportion of correctly predicted instances (all true positives) to the total number of instances. The formula is:

$$\text{Accuracy} = \frac{a + e + i}{a + b + c + d + e + f + g + h + i}$$

   It includes all correct predictions across all classes (the diagonal of the matrix) divided by the total number of instances.

Confusion Matrix

- For Binary Classification

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision, Recall, F1 Score

- We define the following two measures w.r.t. the given target class

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$Precision = \frac{|TP|}{\#\text{ Positive Predictions}} = \frac{|TP|}{|TP| + |FP|}$$

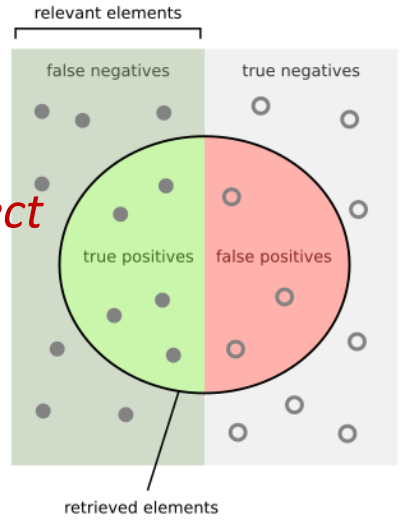
$$Recall = \frac{|TP|}{\#\text{Actual positives}} = \frac{|TP|}{|TP| + |FN|}$$

Predict everything as positive

- *Recall=1, Precision is low*

Predict only 1 positive and it's correct

- *Precision=1, Recall is low*

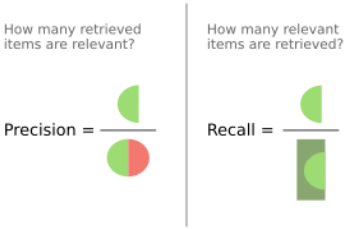


- There is a trade-off between precision and recall

$$F1\ Score = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\mathrm{F}_\beta\ Score = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \text{Precision} + \text{Recall}}$$

Harmonic Mean



Precision, Recall, F1 Score

- Precision: The classifier predict class C; how often is this decision correct?
- Recall: The correct class is C; how often does the classifier predict class C correctly?
- F-measure somewhat combines recall and precision using harmonic mean

Exercise

- Given a classifier f and a test set
  - $X_{test} = \{(x_1, 0), (x_2, 0), (x_3, 1), (x_4, 1)\}$
  - The prediction of f is

	x_1	x_2	x_3	x_4
Actual label	0	0	1	1
prediction	0	1	1	1

- Classification accuracy: 0.75
  - x_1, x_3, x_4 were correctly classified
- Classification error: 1-0.75=0.25
- Precision: 2/3=67%
- Recall: 100%
- F1: 0.8

	Predicted as positive	Predicted as negative
Actual positive	x_3, x_4	NULL
Actual negative	x_2	x_1

Question #7

Classification

- Given the training data in the following table, predict the class of the new data point with Naïve Bayes Classifier.

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	yes	no	no	?

7 Mammals
13 Non-mammals
4 Attributes
20 # of Classes

A: attributes
M: mammals
N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{1}{7} \times \frac{5}{7} \times \frac{2}{7} = 0.025$$

$$P(A|N) = \frac{1}{13} \times \frac{3}{13} \times \frac{10}{13} \times \frac{4}{13} = 0.004$$

$$P(A|M)P(M) = 0.025 \times \frac{7}{20} = 0.00875$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0026$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Summary of Naïve Bayes Classifier

- Easy to implement, good results obtained in most cases
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques
  - E.g., Bayesian Networks

Naïve Bayes Classifier on Example Data

- Given a Test Record $X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- P(X | Yes) =
  P(Refund = No | Yes) x
  P(Divorced | Yes) x
  P(Income = 120K | Yes)

- P(X | No) =
  P(Refund = No | No) x
  P(Divorced | No) x
  P(Income = 120K | No)

Example of Naïve Bayes Classifier

- Given a Test Record $X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$

Naïve Bayes Classifier:

P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/7
P(Marital Status = Divorced | No) = 1/7
P(Marital Status = Married | No) = 4/7
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0

For Taxable Income:
If class = No: sample mean = 110, sample variance = 2975
If class = Yes: sample mean = 90, sample variance = 25

- P(X | No) = P(Refund=No | No)
  × P(Divorced | No)
  × P(Income=120K | No)
  = 4/7 × 1/7 × 0.0072 = 0.0006

- P(X | Yes) = P(Refund=No | Yes)
  × P(Divorced | Yes)
  × P(Income=120K | Yes)
  = 1 × 1/3 × 1.2 × 10^{-9} = 4 × 10^{-10}

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X) => Class = No

Example of Naïve Bayes Classifier

- Given a Test Record $X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$

Naïve Bayes Classifier:

P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/7
P(Marital Status = Divorced | No) = 1/7
P(Marital Status = Married | No) = 4/7
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0

For Taxable Income:
If class = No: sample mean = 110, sample variance = 2975
If class = Yes: sample mean = 90, sample variance = 25

- P(Yes) = 3/10
  P(No) = 7/10
- P(Yes | Divorced) = 1/3 x 3/10 / P(Divorced)
  P(No | Divorced) = 1/7 x 7/10 / P(Divorced)
- P(Yes | Refund = No, Divorced)
  = 1 x 1/3 x 3/10 / P(Divorced, Refund = No)
  P(No | Refund = No, Divorced)
  = 4/7 x 1/7 x 7/10 / P(Divorced, Refund = No)
- P(Yes | Married) = 0 x 3/10 / P(Married)
  P(No | Married) = 4/7 x 7/10 / P(Married)

Issues with Naïve Bayes Classifier

- Consider the table with Tid=7 deleted

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
~~7~~	~~Yes~~	~~Divorced~~	~~220K~~	~~No~~
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naïve Bayes Classifier:

P(Refund = Yes | No) = 2/6
P(Refund = No | No) = 4/6
→ P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/6
→ P(Marital Status = Divorced | No) = 0
P(Marital Status = Married | No) = 4/6
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0/3
For Taxable Income:
If class = No: sample mean = 91, sample variance = 685
If class = Yes: sample mean = 90, sample variance = 25

Given X = (Refund = Yes, Divorced, 120K)

P(X | No) = 2/6 X 0 X 0.0083 = 0

P(X | Yes) = 0 X 1/3 X 1.2 X 10^{-9} = 0

→ Naïve Bayes will not be able to classify X as Yes or No!

In the Naïve Bayes Classifier example provided, the goal is to predict whether a person will "Evade" taxes based on the attributes: Refund, Marital Status, and Taxable Income. The classifier calculates the posterior probabilities $P(X|Yes)$ and $P(X|No)$, where X is the given test record $(Refund = No, Divorced, Income = 120K)$.

Step-by-step explanation of the calculations:

1. **Prior Probability Calculation**:
   The class probabilities $P(Yes)$ and $P(No)$ are based on the number of samples classified as "Yes" or "No" in the dataset.

$$P(Yes) = \frac{3}{10}, \quad P(No) = \frac{7}{10}$$

2. **Calculating Likelihood for $P(X|No)$**:

   For $P(X|No)$, we need to calculate the likelihood of the attributes given the class "No" and multiply them together:

   - $P(Refund = No|No)$:
     This is the probability that the refund is "No" given that the person did not evade taxes.

$$P(Refund = No|No) = \frac{4}{7}$$

   - $P(MaritalStatus = Divorced|No)$:
     This is the probability that the marital status is "Divorced" given that the person did not evade taxes.

$$P(MaritalStatus = Divorced|No) = \frac{1}{7}$$

   - $P(Income = 120K|No)$:
     Since income is continuous, we use a Gaussian distribution to estimate this probability. The formula for the probability is:

$$P(Income = 120K|No) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(120-\mu)^2}{2\sigma^2}\right)$$

     where the mean μ and variance σ^2 for income when the class is "No" are given as:

$$\mu = 110, \quad \sigma^2 = 2975$$

     Substituting the values:

$$P(Income = 120K|No) \approx 0.0072$$

Now multiply all these probabilities:

$$P(X|No) = \frac{4}{7} \times \frac{1}{7} \times 0.0072 \approx 0.0006$$

3. **Calculating Likelihood for $P(X|Yes)$:**

   For $P(X|Yes)$, we calculate the likelihood of the attributes given the class "Yes":

   - $P(Refund = No|Yes)$:

     $$P(Refund = No|Yes) = 1 \quad (\text{since all "Yes" cases have Refund} = \text{No})$$

   - $P(MaritalStatus = Divorced|Yes)$:

     $$P(MaritalStatus = Divorced|Yes) = \frac{1}{3}$$

   - $P(Income = 120K|Yes)$:
     Using the Gaussian distribution formula for income, with:

     $$\mu = 90, \quad \sigma^2 = 25$$

     Substituting the values:

     $$P(Income = 120K|Yes) \approx 1.2 \times 10^{-9}$$

   Now multiply all these probabilities:

   $$P(X|Yes) = 1 \times \frac{1}{3} \times 1.2 \times 10^{-9} \approx 4 \times 10^{-10}$$

4. **Calculating Posterior Probability:**

   To calculate the final probability for each class, multiply the likelihoods by the prior probabilities:

   - For "No":

     $$P(X|No)P(No) = 0.0006 \times \frac{7}{10} = 0.00042$$

   - For "Yes":

     $$P(X|Yes)P(Yes) = 4 \times 10^{-10} \times \frac{3}{10} = 1.2 \times 10^{-10}$$

5. **Comparing the Probabilities:**

   Since $P(X|No)P(No) > P(X|Yes)P(Yes)$, the Naïve Bayes classifier predicts the class as "No" (i.e., the person will **not evade** taxes).

Question #8

Classification

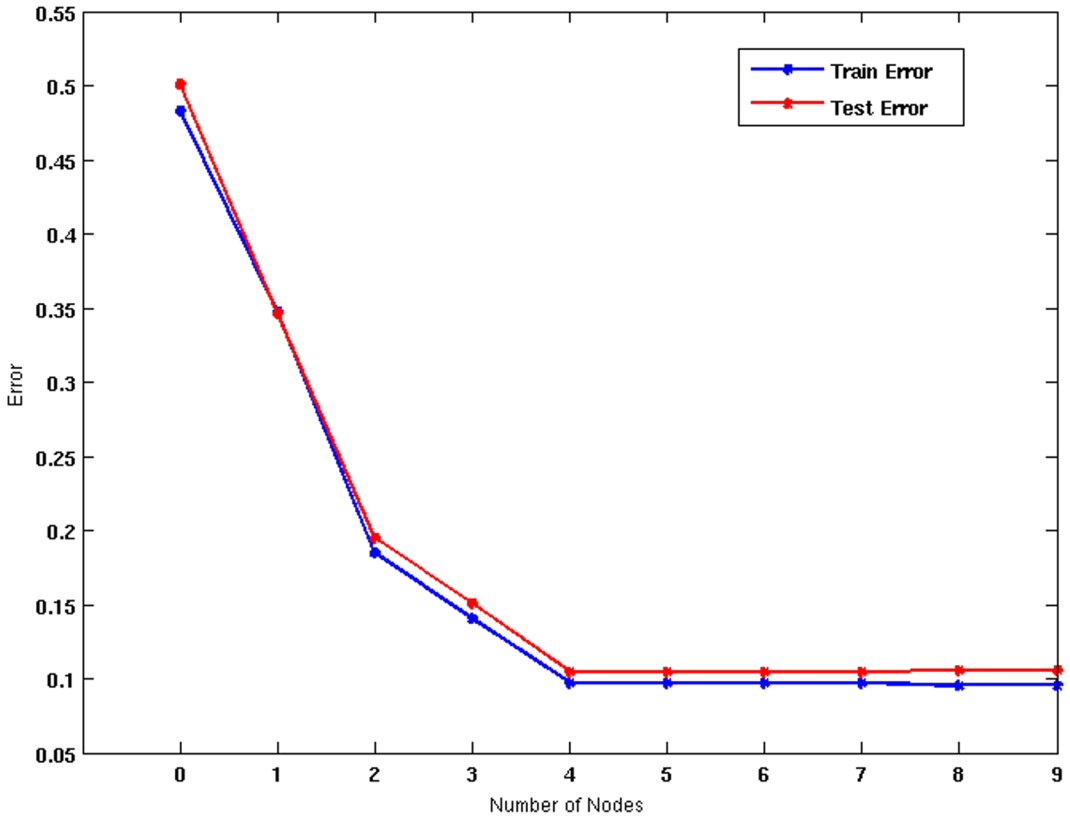
- Please describe what are underfitting and overfitting, and what methods can be used to avoid overfitting.

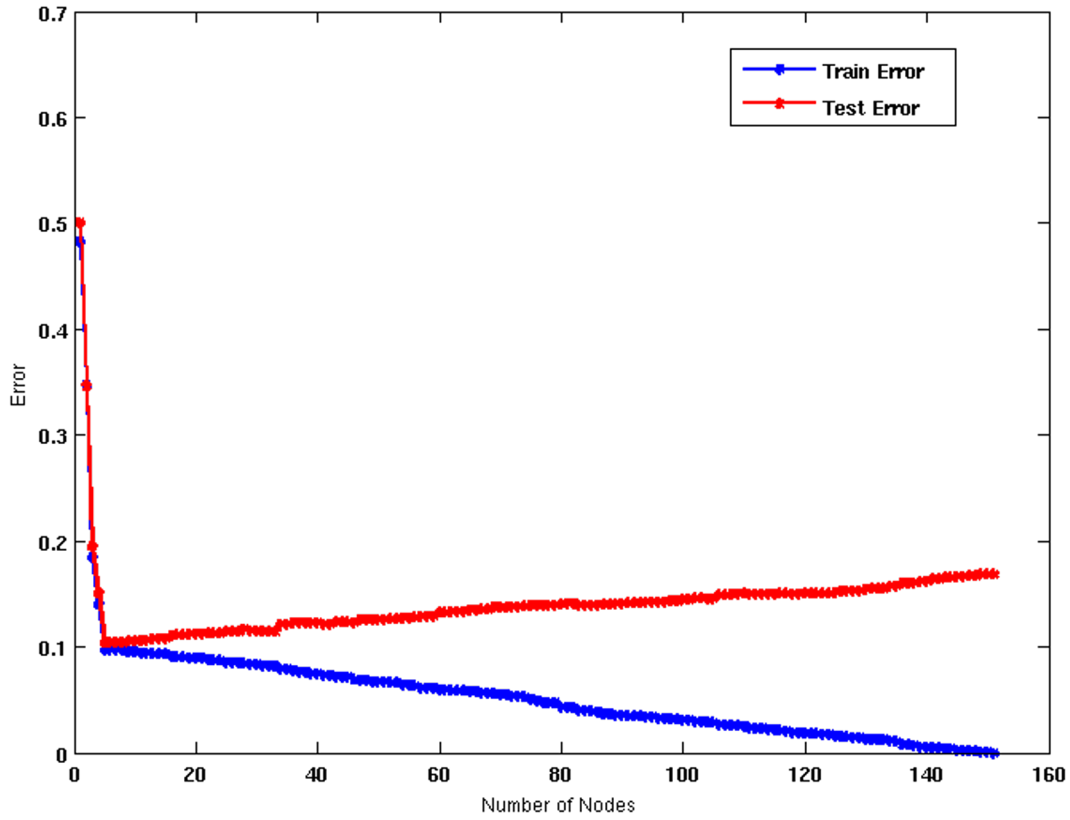
Underfitting: a scenario where a model is unable to capture the relationship between the input and output variables accurately, generating a high error rate on both the training set and test set data.

Overfitting: a scenario where a model fits the training set too much, even fits the noise in the training set well, making the model unable to correctly predict new data, resulting in small training error but large test error.

Method: enlarge training set size; use simpler models; early stopping training algorithm; post-pruning (for decision trees).

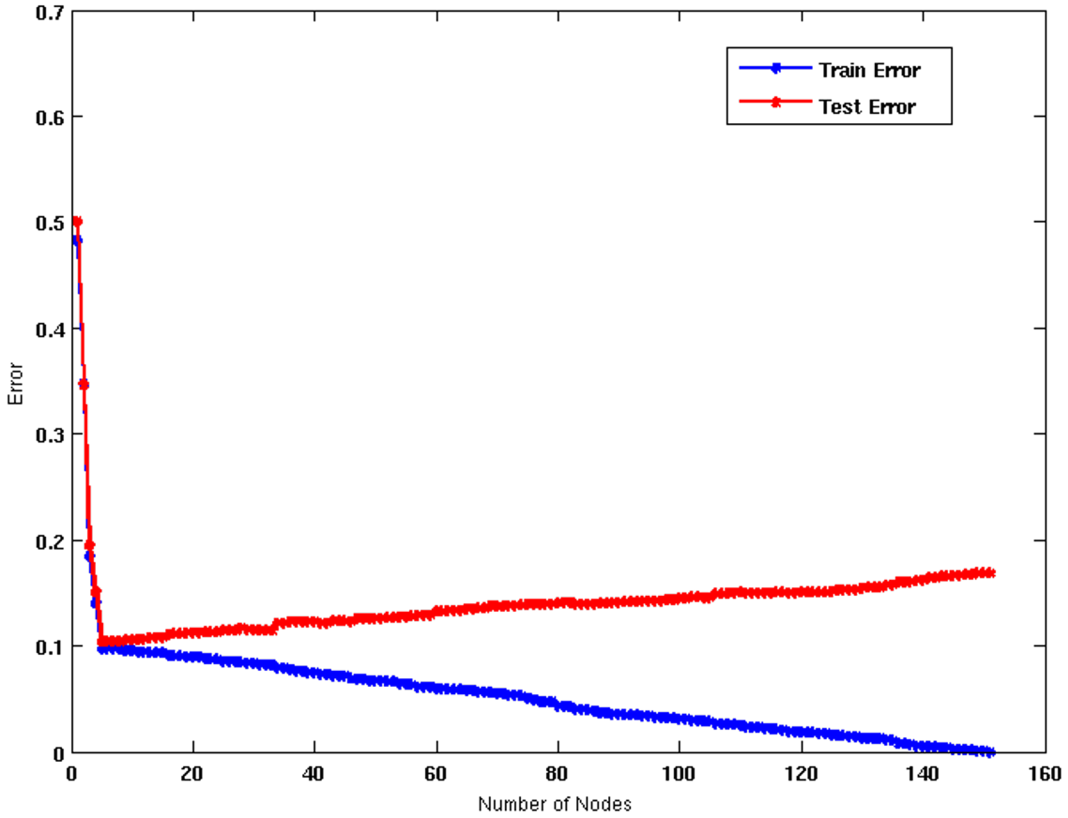
Model Overfitting

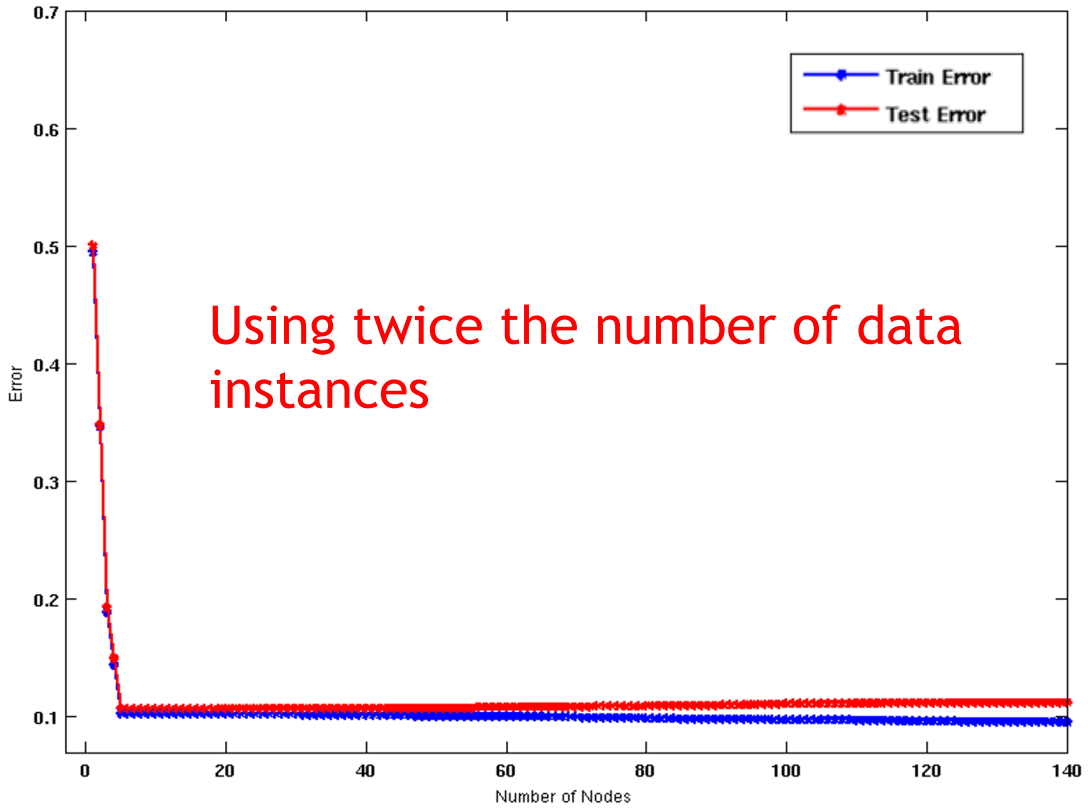




- **Underfitting**: when model is too simple, both training and test errors are large
- **Overfitting**: when model is too complex, training error is small but test error is large

Model Overfitting





Using twice the number of data instances

- If training data is under-representative, testing errors increase and training errors decrease on increasing number of nodes
- Increasing the size of training data reduces the difference between training and testing errors at a given number of nodes