Final Review

** This is the review of some post-midterm topics.
This review is NOT exhaustive.
You are responsible for covering the entire course.
The final will be comprehensive **

Anurag Nagar

Machine Learning Class

Topics Covered

List of topics covered post midterm:

- Logistic Regression
- Model Evaluation
- Instance Based Learning (k-Nearest Neighbor)
- Ensemble Learning
- Unsupervised Learning, PCA
- Bayes Net
- Hidden Markov Models (HMM)

Outline

Logistic Regression

Creates a probabilistic model for classes i.e. tries to predict $P(Y = y_k)$ where y_k could be 0 or 1 for a binary dataset.

$$p = P(Y = 1) = \frac{1}{1 + e^{-z}} \text{ where } z = w_0 + \sum_{i=1}^{n} w_i x_i \text{ and}$$

$$1 - p = P(Y = 0) = \frac{e^{-z}}{1 + e^{-z}}$$

$$ln(\frac{p}{1-p}) = z = w_0 + \sum_{i=1}^{n} w_i x_i$$

so log of odds ratio is a linear function of x's

Logistic Regression

As always,

$$W = \begin{bmatrix} w_0 \\ w_1 \\ \dots \\ w_n \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} x_0 \\ x_1 \\ \dots \\ x_n \end{bmatrix}$$

where $x_0 = 1$

In some questions, you are given W and X and are asked if the data is more likely to belong to class 1 or class 0. You may also be asked to find out the dividing boundary. You need to check if

$$P(Y = 1) > P(Y = 1)$$

for given values.

Logistic Regression

Note: z is a dot product of vectors W and X

$$z = W \cdot X$$

Assign to class 1 if:

$$\frac{p}{1-p} > 1 \quad \text{or}$$

$$z > 0 \quad \text{or}$$

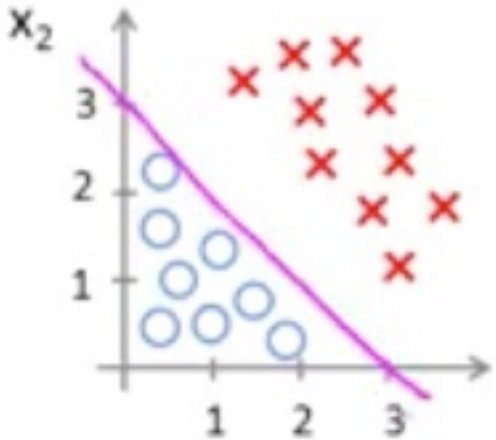
$$w_0 + \sum_{i=1}^{n} w_i x_i > 0$$

Logistic Regression

If

$$W = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

Class is 1 if



$$w_0 + \sum_{i=1}^{n} w_i x_i > 0$$

$$-3 + x_1 + x_2 > 0$$

Red is class 1

Question 1

Difference between generative and discriminative classifiers?

Question 1

Difference between generative and discriminative classifiers?

Generative classifiers estimate P(Y|X) from the training data using Bayes rule. In other words, they estimate joint probability of y and x from the data.
Discriminative classifiers make a model of P(Y|X)

Question 2

If there are n Booleans $X_1, X_2, \ldots X_n$, and they are all dependent on each other, how many parameters would be needed to describe their joint probability $P(X_1, X_2, \ldots X_n)$

Question 2

If there are n Booleans $X_1, X_2, \ldots X_n$, and they are all dependent on each other, how many parameters would be needed to describe their joint probability $P(X_1, X_2, \ldots X_n)$

$2^n - 1$

Question 3

If there are n Booleans $X_1, X_2, \ldots X_n$, and they are all independent of each other, how many parameters would be needed to describe their joint probability $P(X_1, X_2, \ldots X_n)$

Question 3

If there are n Booleans $X_1, X_2, \ldots X_n$, and they are all independent of each other, how many parameters would be needed to describe their joint probability $P(X_1, X_2, \ldots X_n)$

n

Question 4

If there are 2 Boolean classes Y_1, Y_2 Booleans, and each class has n Boolean features. We assume inter-class feature independence i.e. feature of a class depend on each other, but are independent of features in the other class. How many parameters are needed to describe the system.

Question 4

If there are 2 Boolean classes Y_1, Y_2 Booleans, and each class has n Boolean features. We assume inter-class feature independence i.e. feature of a class depend on each other, but are independent of features in the other class. How many parameters are needed to describe the system.

$2 \times (2^n - 1)$

Question 5

If there are 2 Boolean classes Y_1, Y_2 each consisting of n Boolean features, and each class is modeled using the traditional logistic regression function

$$P(Y = Y_i|(X_1, X_2, \ldots, X_n)) = \frac{1}{1 + exp(w_0 + \sum_{j=1}^{n} w_j X_j)}$$

How many total parameters are needed to describe the system?

Question 5

If there are 2 Boolean classes Y_1, Y_2 each consisting of n Boolean features, and each class is modeled using the traditional logistic regression function

$$P(Y = Y_i|(X_1, X_2, \ldots, X_n)) = \frac{1}{1 + exp(w_0 + \sum_{j=1}^{n} w_j X_j)}$$

How many total parameters are needed to describe the system?

$(n + 1)$

Question 6

Suppose you are playing a game that involves hitting a target. You are given 3 chances, but the games gets tougher every time you miss. The probabilities of hitting in the first, second, and third attempts is 0.6, 0.5, 0.4
Find the odds ratio for this game.

Question 6

Suppose you are playing a game that involves hitting a target. You are given 3 chances, but the games gets tougher every time you miss. The probabilities of hitting in the first, second, and third attempts is 0.6, 0.5, 0.4
Find the odds ratio for this game.

Probability of failure:
$1 - p = (1 - 0.6) \times (1 - 0.5) \times (1 - 0.4) = 0.12$
Probability of success: $p = 0.88$
Odds Rato $= \frac{0.88}{0.12} = 7.33$

Question 7

Logistic Regression is what type of model?
Linear or Non-Linear

Question 7

Logistic Regression is what type of model?
Linear or Non-Linear

Linear (Proved in class)

Question 8

Suppose the logistic regression for binary class $Y = 0$ and $Y = 1$ and two feature X_1, X_2 problem is defined as below:

$$P(Y = 1|(X_1, X_2)) = \frac{1}{1 + exp(w_0 + \sum_{j=1}^{2} w_j X_j)}$$

The values of w_0, w_1, w_2 are -1, 2, 0 respectively. What will be the separation rule for the two classes.

Question 8

Suppose the logistic regression for binary class $Y = 0$ and $Y = 1$ and two feature X_1, X_2 problem is defined as below:

$$P(Y = 1|(X_1, X_2)) = \frac{1}{1 + exp(w_0 + \sum_{j=1}^{2} w_j X_j)}$$

The values of w_0, w_1, w_2 are -1, 2, 0 respectively. What will be the separation rule for the two classes.

For data to be in class 1:
$P(Y = 1) > P(Y = 0)$ or
$w_0 + w_1 x_1 + w_2 x_2 > 0$ or
$1 + 2 \times x_1 + 0 \times x_2 > 0$ or
$-1 + 2x_1 > 0$ or
$x_1 > \frac{1}{2}$

Question 9

For estimating best parameters of logistic regression, which technique is used:

Question 9

For estimating best parameters of logistic regression, which technique is used:

MLE + Gradient Ascent

Logistic Regression Question

Consider the entropy function for a binary class problem:

$$H = p\ log_2(p) + (1-p)\ log_2(1-p)$$

where p is probability of class 1. A plot of H vs p would be what type of curve?

1. Concave
2. Linear
3. Convex
4. Polynomial of degree 4

Logistic Regression Question

Consider the entropy function for a binary class problem:

$$H = p\ log_2(p) + (1-p)\ log_2(1-p)$$

where p is probability of class 1. A plot of H vs p would be what type of curve?

1. Concave
2. Linear
3. Convex
4. Polynomial of degree 4

Outline

Model Evaluation

Consider the confusion matrix shown below:

	Predicted = YES	Predicted = NO
Actual = YES	100	40
Actual = NO	10	50

Calculate:

- Accuracy
- Recall
- Precision
- F-Statistic

Model Evaluation

Consider the confusion matrix shown below:

	Predicted = YES	Predicted = NO
Actual = YES	100	40
Actual = NO	10	50

Calculate:

- Accuracy $= \frac{150}{200}$
- Precision (p) $= \frac{100}{110}$
- Recall (r) $= \frac{100}{140}$
- F-Statistic $= \frac{2rp}{r+p} = 0.80$

Model Evaluation

Consider the following confusion matrix.

Number of cases: **100,000**

Actual State	Predicted Negative	Predicted Positive
Negative	TN **97750**	FP **150**
Positive	FN **330**	TP **1770**

Compute precision, recall, F-statistic

Model Evaluation

Consider the following confusion matrix.

Number of cases: **100,000**

Actual State	Predicted Negative	Predicted Positive
Negative	TN **97750**	FP **150**
Positive	FN **330**	TP **1770**

Compute precision, recall, F-statistic

Should be easy

ROC Curve

Remember the definitions:

- True Positive Rate (TPR) $= \frac{TP}{TP + FN}$
- False Positive Rate (FPR) $= \frac{FP}{FP + TN}$
- Receiver Operative Characteristic (ROC) curve is a plot of TPR vs FPR

Model Evaluation Question

I am testing a model that predicts the probability of a person having disease A. I have obtained the following results:

Person	Predicted probability	True Value
1	0.95	1
2	0.80	1
3	0.60	1
4	0.45	1
5	0.15	0

Calculate area under ROC curve by setting the positive threshold i.e. probability equal to or greater than which prediction is labeled as positive :

- 0.0
- 0.25
- 0.50
- 0.75
- 1.0

What according to you should be the best threshold for predicted probability?

Model Evaluation Question

Threshold	0.0	0.25	0.5	0.75	1.0
TP	4	4	3	2	0
FP	1	0	0	0	0
TN	0	1	1	1	1
FN	0	0	1	2	4
TPR	1.0	1.0	0.75	0.50	0.0
FPR	1.0	0.0	0.0	0.0	0.0

Area under ROC = 100.0%

Best threshold = 0.25

Model Evaluation Question

I am testing a model that predicts the probability of a person having disease A. I have obtained the following results:

Person	Predicted probability	True Value
1	0.80	1
2	0.60	0
3	0.40	1
4	0.20	0

Calculate area under ROC curve by setting the positive threshold i.e. probability equal to or greater than which prediction is labeled as positive :

- 0.0
- 0.2
- 0.4
- 0.6
- 0.8
- 1.0

What according to you should be the best threshold for predicted probability?

Model Evaluation Question

I am testing a model that predicts the probability of a person having disease A. I have obtained the following results:

Person	Predicted probability	True Value
1	0.80	1
2	0.60	0
3	0.40	1
4	0.20	0

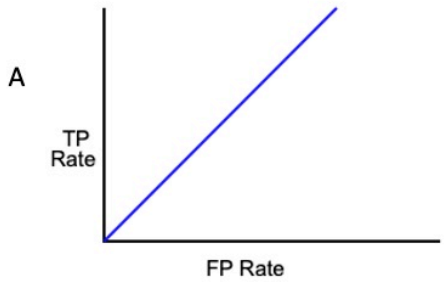
Calculate area under ROC curve by setting the positive threshold i.e. probability equal to or greater than which prediction is labeled as positive :

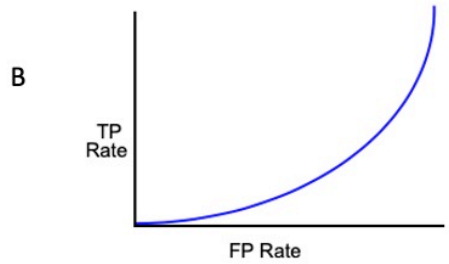
- 0.0
- 0.2
- 0.4
- 0.6
- 0.8
- 1.0

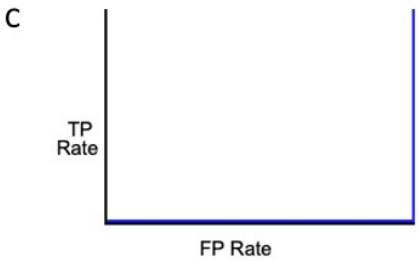
What according to you should be the best threshold for predicted probability? Try yourself

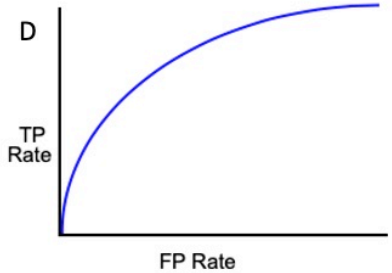
Model Evaluation

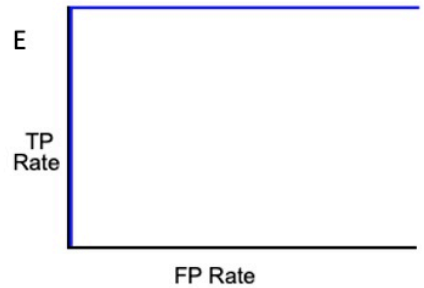
Consider the following ROC curves.







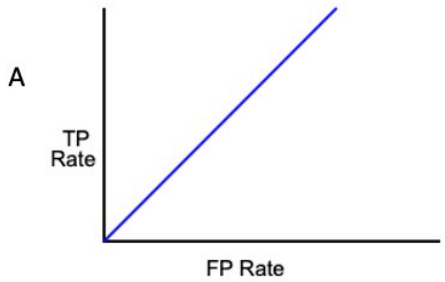


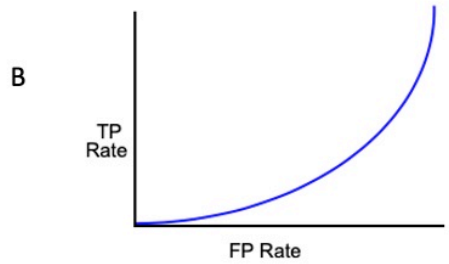


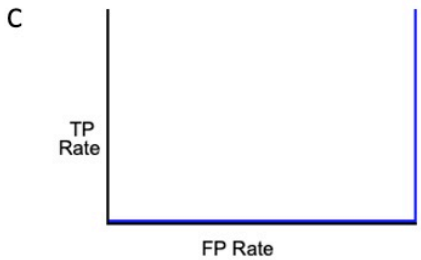
Arrange the curves from best to worst

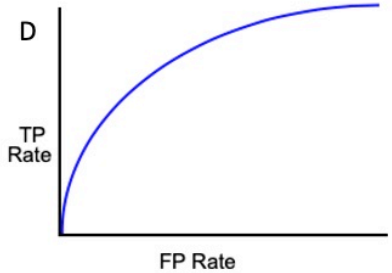
Model Evaluation

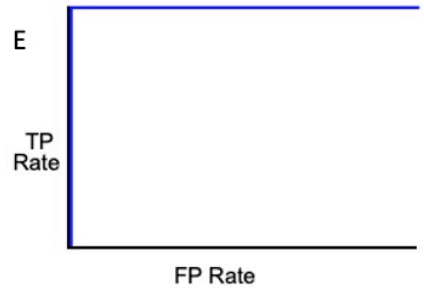
Consider the following ROC curves.











Arrange the curves from best to worst

E, D, A, B, C

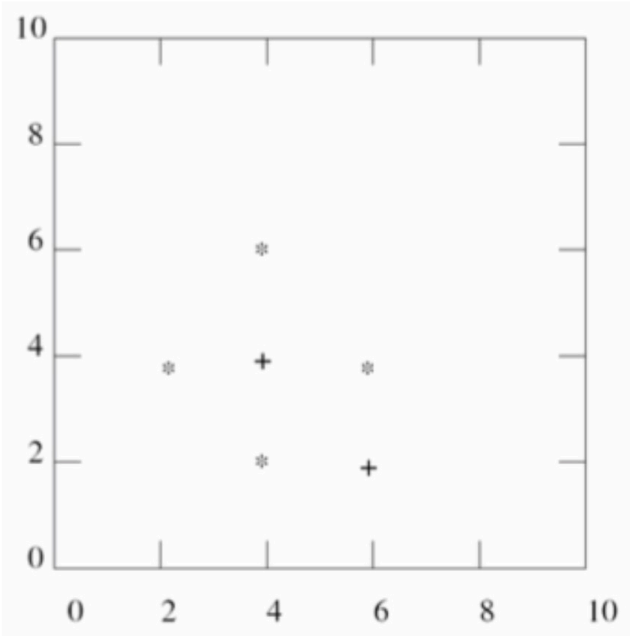
Outline

Instance Based Learning

- In k-NN, the parameter k is a hyperparameter not a model parameter. Understand the differences.
- Understand training and test phases.
- Practice questions on predictions using k-NN
- Bias and Variance versus k in k-NN

Instance Based Learning

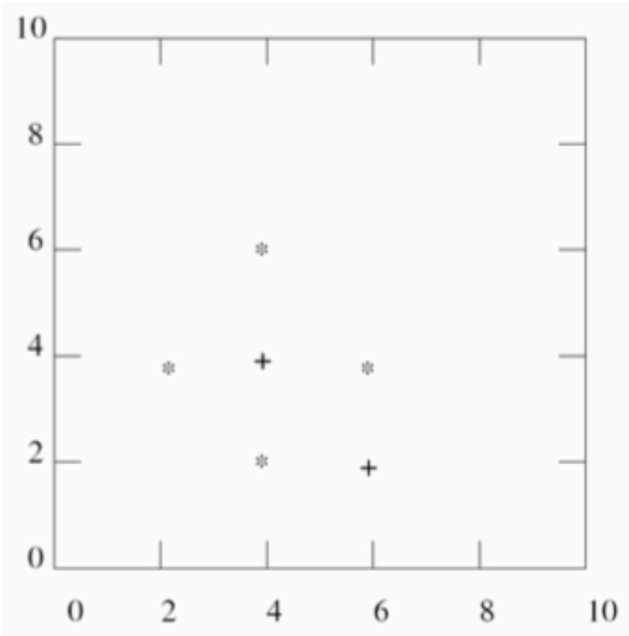
Consider the dataset shown below:



What will be the error of 1-NN and 3-NN applied to this dataset?

Instance Based Learning

Consider the dataset shown below:



What will be the error of 1-NN and 3-NN applied to this dataset?

1-NN $= 6/6$, 3-NN $= 4/6$

Instance Based Learning

Consider the dataset shown below:

	Solid	Temp	Force Applied	Class
1	Yes	0	0.406	No
2	No	1	0.25	No
3	No	0	0.063	No
4	Yes	1	0.375	No
5	No	0.5	0.219	Yes
6	No	1	0	No
7	Yes	0.5	1	No
8	No	0	0.156	Yes
9	No	1	0.094	No
10	No	0	0.188	Yes

Using 1-NN, what will be the prediction for item (No, 0, 0.375)? You can assume Manhattan distance for each of the attributes. For the first attribute, you can assume Yes $= 1$ and No $= 0$

Instance Based Learning

Consider the dataset shown below:

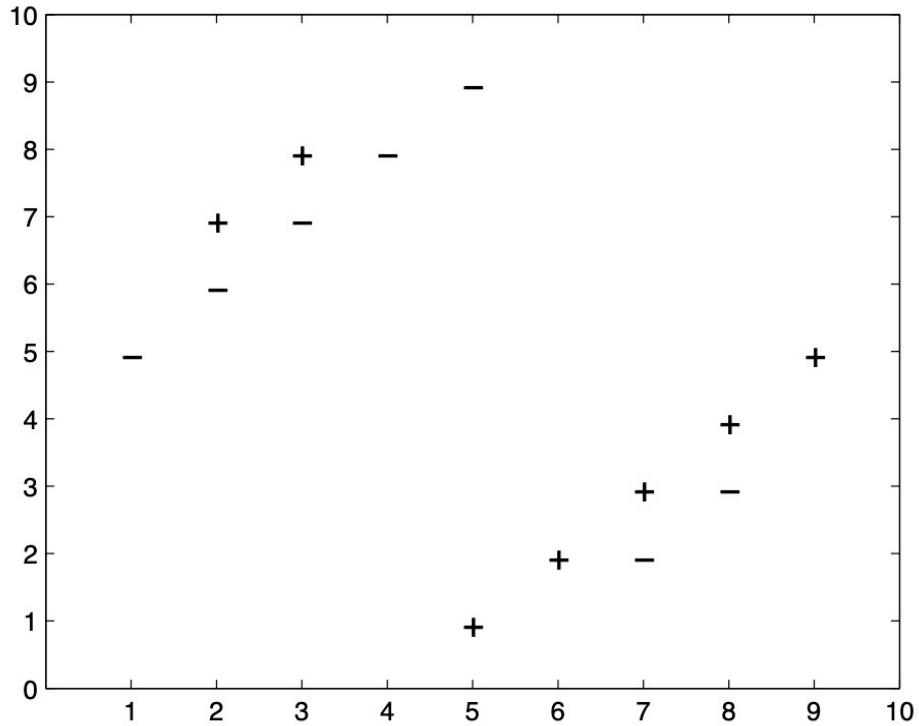
	Solid	Temp	Force Applied	Class
1	Yes	0	0.406	No
2	No	1	0.25	No
3	No	0	0.063	No
4	Yes	1	0.375	No
5	No	0.5	0.219	Yes
6	No	1	0	No
7	Yes	0.5	1	No
8	No	0	0.156	Yes
9	No	1	0.094	No
10	No	0	0.188	Yes

Using 1-NN, what will be the prediction for item (No, 0, 0.375)? You can assume Manhattan distance for each of the attributes. For the first attribute, you can assume Yes $= 1$ and No $= 0$

Class = Yes

Instance Based Learning

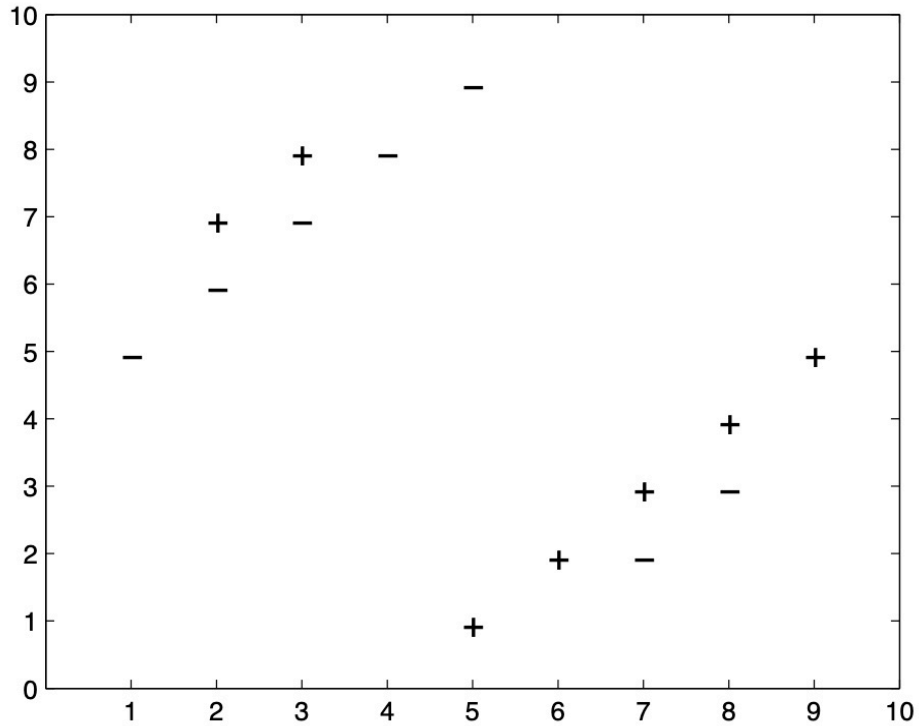
Consider the visual dataset shown below:



What value of k in k-NN will give the best and the worst training error? Can you obtain 0 training error for any value of k?

Instance Based Learning

Consider the visual dataset shown below:



What value of k in k-NN will give the best and the worst training error? Can you obtain 0 training error for any value of k?

Try various values of k, and investigate

Outline

Ensemble Methods

Suppose I create a model for regression, such as $y = w_0 + \sum_{i=1}^{n} w_i x^i$. I can fit the *training data* nicely by making a complex polynomial i.e. increasing the value of n. What effect will it cause to bias and variance

Ensemble Methods

Suppose I create a model for regression, such as $y = w_0 + \sum_{i=1}^{n} w_i x^i$. I can fit the *training data* nicely by making a complex polynomial i.e. increasing the value of n. What effect will it cause to bias and variance

Decrease bias and increase variance

Ensemble Methods

Suppose I create a model for regression, such as $y = w_0 + \sum_{i=1}^{n} w_i x^i$. I can fit the *training data* nicely by making a complex polynomial i.e. increasing the value of n. How can I prevent the model from becoming too complex?

Ensemble Methods

Suppose I create a model for regression, such as $y = w_0 + \sum_{i=1}^{n} w_i x^i$. I can fit the *training data* nicely by making a complex polynomial i.e. increasing the value of n. How can I prevent the model from becoming too complex?

Add a penalty term called regularization:
Total Error $= \sum_{i=1}^{M} (t_i - y_i)^2 + \lambda \sum_{i=1}^{n} w_i^2$
M is the number of data points and n is the degree of polynomial used.

https://towardsdatascience.com/regularization-an-important-concept-in-machine-learning-5891628907ea

https://datanice.github.io/machine-learning-101-what-is-regularization-interactive.html

https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a

Ensemble Methods

How do you control bias and variance in the following models:

- Linear Regression
- Neural Networks
- Decision Trees
- Logistic Regression
- k-Nearest Neighbors

Ensemble Methods

How do you control bias and variance in the following models:

- Linear Regression
- Neural Networks
- Decision Trees
- Logistic Regression
- k-Nearest Neighbors

Do your research and find answers.

Ensemble Methods

How does an ensemble (set of classifiers working together) prevent overfitting and still give a good accuracy?

Ensemble Methods

How does an ensemble (set of classifiers working together) prevent overfitting and still give a good accuracy?

Averaging of models gives a more stable model overall i.e. a model with low variance.
It's like taking the opinion of multiple people before taking a decision.

https://towardsdatascience.com/simple-guide-for-ensemble-learning-methods-d87cc68705a2

https://machinelearningmastery.com/ensemble-methods-for-deep-learning-neural-networks/

Ensemble Methods

Practice examples of Adaboost.
Like the ones solved in class, and

http://ais.informatik.uni-freiburg.de/teaching/ws11/robotics2/pdfs/rob2-18-adaboost.pdf

http://www.authorstream.com/Presentation/aSGuest79199-727683-animated-adaboost-example/

Conceptual knowledge of Adaboost, such as:

Can Adaboost using stumps (i.e. decision tree of depth 1) always give 0 *training* error ?

Will Adaboost always converge?

Random Forest

In Random forest you can generate large number of trees (say $T_1, T_2 \ldots .. T_n$) and then aggregate the results of these trees. Which of the following is true about individual(T_k) tree in Random Forest?

1. Individual tree is built on a subset of the features
2. Individual tree is built on all the features
3. Individual tree is built on a subset of observations
4. Individual tree is built on full set of observations

Random Forest

In Random forest you can generate large number of trees (say $T_1, T_2 \ldots .. T_n$) and then aggregate the results of these trees. Which of the following is true about individual(T_k) tree in Random Forest?

1. Individual tree is built on a subset of the features
2. Individual tree is built on all the features
3. Individual tree is built on a subset of observations
4. Individual tree is built on full set of observations

Outline

K-means

We would like to perform K-means clustering on the four point shown below:
x1 = (1, 1)
x2 = (2, 2)
x3 = (6, 6)
x4 = (7, 7)
The initial centers are x1 and x2 for clusters C1 and C2 respectively. We will use Manhattan distance.

For the first iteration, which points will be assigned to C1 and C2?

New centers after first iteration:

K-means

We would like to perform K-means clustering on the four point shown below:
x1 = (1, 1)
x2 = (2, 2)
x3 = (6, 6)
x4 = (7, 7)
The initial centers are x1 and x2 for clusters C1 and C2 respectively. We will use Manhattan distance.

For the first iteration, which points will be assigned to C1 and C2?
x1 to C1, and x2, x3, x4 to C2
New centers after first iteration:
C1 = (1, 1), C2 = (5, 5)

K-means

We would like to perform K-means clustering on the four point shown below:
x1 = (1, 1)
x2 = (2, 2)
x3 = (6, 6)
x4 = (7, 7)
The initial centers are x1 and x2 for clusters C1 and C2 respectively. We will use Manhattan distance.

For the second iteration, which points will be assigned to C1 and C2?
New centers after second iteration:

K-means

We would like to perform K-means clustering on the four point shown below:
x1 = (1, 1)
x2 = (2, 2)
x3 = (6, 6)
x4 = (7, 7)
The initial centers are x1 and x2 for clusters C1 and C2 respectively. We will use Manhattan distance.

For the second iteration, which points will be assigned to C1 and C2?
x1, x2 to C1, and x3, x4 to C2
New centers after second iteration:
C1 = (1.5, 1.5), C2 = (6.5, 6.5)

K-means

We would like to perform K-means clustering on the four point shown below:
x1 = (1, 1)
x2 = (2, 2)
x3 = (6, 6)
x4 = (7, 7)
The initial centers are x1 and x2 for clusters C1 and C2 respectively. We will use Manhattan distance.

For the third iteration, which points will be assigned to C1 and C2?

New centers after third iteration:

K-means

We would like to perform K-means clustering on the four point shown below:
x1 = (1, 1)
x2 = (2, 2)
x3 = (6, 6)
x4 = (7, 7)
The initial centers are x1 and x2 for clusters C1 and C2 respectively. We will use Manhattan distance.

For the third iteration, which points will be assigned to C1 and C2?
x1, x2 to C1, and x3, x4 to C2
New centers after third iteration:
C1 = (1.5, 1.5), C2 = (6.5, 6.5) No change, so can stop

K-means

We would like to perform K-means clustering on the four point shown below:
x1 = (1, 1)
x2 = (2, 2)
x3 = (6, 6)
x4 = (7, 7)
The initial centers are x1 and x2 for clusters C1 and C2 respectively. We will use Manhattan distance.

What will be the value of Within Cluster Sum of Squares (WCSS) after the end of clustering?

K-means

We would like to perform K-means clustering on the four point shown below:
x1 = (1, 1)
x2 = (2, 2)
x3 = (6, 6)
x4 = (7, 7)
The initial centers are x1 and x2 for clusters C1 and C2 respectively. We will use Manhattan distance.

What will be the value of Within Cluster Sum of Squares (WCSS) after the end of clustering?
Cluster 1: $1^2 + 1^2$, Cluster 2: $1^2 + 1^2$
Total = 4.0

Unsupervised Learning

Suppose I have the following dataset:

Point	x_1	x_2
A	1	1
B	1	0
C	3	3
D	2	4
E	3	5

Perform k-means clustering with k $= 2$, with A and C as initial centers. Use Manhattan distance.

Unsupervised Learning

First Iteration

Starting distance:

from A: A = 0, B = 1, C = 4, D = 4, E = 6

from C: A = 4, B = 5, C = 0, D = 2, E = 2

Update Centers:

C1 = (1, 0.5)

C2 = (2.67, 4)

Second Iteration

Starting distance:

from C1: A = 0.5, B = 0.5, C = 4.5, D = 4.5, E = 6.5

from C2: A = 4.67, B = 5.67, C = 0.67, D = 0.67, E = 1.33

Update Centers:

C1 = (1, 0.5)

C2 = (2.67, 4)

No change in assignment or centers, so can stop

Unsupervised Learning

Suppose I have the following dataset:

Point	x_1	x_2
A	1	1
B	1	0
C	3	3
D	2	4
E	3	5

Perform single link hierarchical clustering and draw the final dendrogram. Use Manhattan distance.

Unsupervised Learning

Suppose I have the following dataset:

Point	x_1	x_2
A	1	1
B	1	0
C	3	3
D	2	4
E	3	5

Perform single link hierarchical clustering and draw the final dendrogram. Use Manhattan distance.

Try yourself

PCA

Which of the following are aims of PCA analysis?

1. To convert a set of correlated attributes of a dataset into features that are orthonormal to each other.
2. To discover a set of principal components that explain variance in data.
3. To rename columns.
4. To remove duplicate rows.

PCA

Which of the following are aims of PCA analysis?

1. To convert a set of correlated attributes of a dataset into features that are orthonormal to each other.
2. To discover a set of principal components that explain variance in data.
3. To rename columns.
4. To remove duplicate rows.

PCA

Which of the following is true about the first principal component?

1. It is a linear combination of other principal components and explains the largest percentage of variance in the dataset.
2. It explains the least variance in the dataset.
3. It is orthogonal to all other principal components and explains the largest percentage of variance in the dataset.
4. Principal components are randomly arranged.

PCA

Which of the following is true about the first principal component?

1. It is a linear combination of other principal components and explains the largest percentage of variance in the dataset.
2. It explains the least variance in the dataset.
3. It is orthogonal to all other principal components and explains the largest percentage of variance in the dataset.
4. Principal components are randomly arranged.

PCA

Understand the concept of PCA, explained variance, etc. No numerical calculations will be asked.

Outline

Bayes Net

Suppose I have a Bayes net of 5 nodes, where each node can take 4 possible values i.e. the nodes are not binary. How many parameters would be needed if they were all dependent, and when they were all fully independent.

Bayes Net

Practice more questions involving joint probability, summing out etc, dependency separation (d-separation), etc

Outline

Bayes Net

Practice Forward and Viterbi algorithm questions.