MidTerm Review

Anurag Nagar

Machine Learning Class

Outline

Topics Covered

List of topics covered so far (in no particular order)

- Inductive Learning - Linear Regression
- Information Based Learning - Entropy and Information Gain, Decision Trees
- Probability and naive Bayes Classifier
- Perceptron
- Artificial Neural Network
- SVM

Deep Learning will **not** be included

Outline

Inductive Learning

What is Inductive Learning?

- A dataset is available with following data for each instance:
  target variable Y
  feature vector $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$

Inductive Learning

What is Inductive Learning?

- A dataset is available with following data for each instance:
  target variable Y

  feature vector $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$

- You assume that there is a function $f(X)$ that relates X to Y.

  $$Y = f(X) + \epsilon$$

  where ϵ is the error.

Inductive Learning

What is Inductive Learning?

- A dataset is available with following data for each instance: target variable Y

  feature vector $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$

- You assume that there is a function $f(X)$ that relates X to Y.

$$Y = f(X) + \epsilon$$

  where ϵ is the error.

- For example, Y could be the GPA, and X_1 could be the SAT score, X_2 could be percent attendance, and X_3 could be the hours studied every week.

Inductive Learning

Function Estimation

- Based on the training data, you propose a function $\hat{f}$ that approximates f

Inductive Learning

Function Estimation

- Based on the training data, you propose a function $\hat{f}$ that approximates f
- You would like $\hat{f}$ to be such that it minimizes the mean squared error over all examples i

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{f}(x_i))^2$$

Inductive Learning

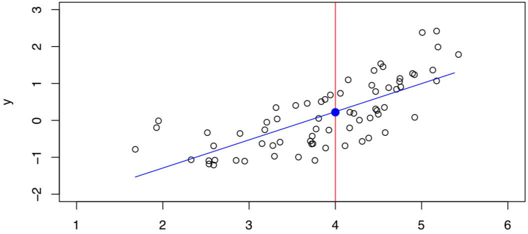


Figure: Linear Model

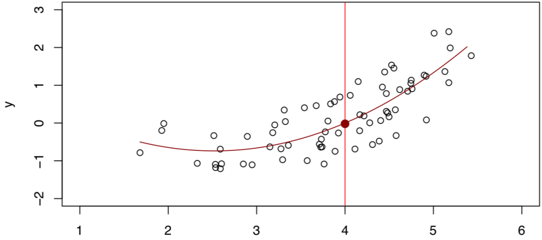
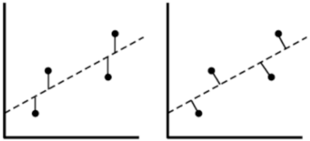


Figure: Quadratic Model

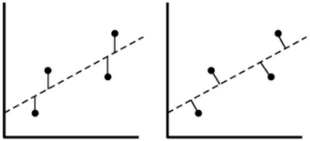
Example 1

In the figures below, the circles represent the data points, and the dashed line represents the function fitted to the line. Which of the following shows the correct value of error for each point?



Example 1

In the figures below, the circles represent the data points, and the dashed line represents the function fitted to the line. Which of the following shows the correct value of error for each point?



The one on the left as it shows the distance between prediction and actual value.

Example 2

For predicting the output variable Y as a function of input variable x, you design a linear model of the following form:

$$Y = a + bx$$

What do a and b represent?

Example 2

For predicting the output variable Y as a function of input variable x, you design a linear model of the following form:

$$Y = a + bx$$

What do a and b represent?

a represents the Y-intercept and b represents the slope.

Example 3

For the following dataset, x represents the independent variable, and y is the output variable.

x	1	10	20
y	1	100	400

You propose a linear regression equation as below:

$$Y = -140 + 30x$$

What will be the value of Mean Squared Error (MSE)?

Example 3

For the following dataset, x represents the independent variable, and y is the output variable.

x	1	10	20
y	1	100	400

You propose a linear regression equation as below:

$$Y = -140 + 30x$$

What will be the value of Mean Squared Error (MSE)?

The predicted values for each of the points would be: -110, 160, and 460. So the MSE would be:

$$MSE = \frac{1}{3}[(1+110)^2 + (100-160)^2 + (400-460)^2] = 6507$$

Outline

Decision Trees

Which of the following is/are types of nodes in a decision tree

1. Root Node
2. Leaf Node
3. Internal Node
4. Jump Node

Decision Trees

Which of the following is/are types of nodes in a decision tree

1. Root Node
2. Leaf Node
3. Internal Node
4. Jump Node

Decision Trees

Which of the following node in a decision tree gives the predicted label?

1. Root Node
2. Leaf Node
3. Internal Node
4. Jump Node

Decision Trees

Which of the following node in a decision tree gives the predicted label?

1. Root Node
2. Leaf Node
3. Internal Node
4. Jump Node

Decision Trees

In a decision tree, which node contains all of the input data?

1. Root Node
2. Leaf Node
3. Internal Node
4. Jump Node

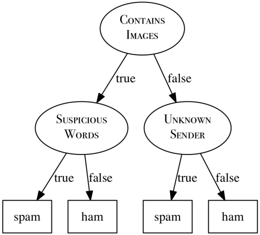
Decision Trees

In a decision tree, which node contains all of the input data?

1. Root Node
2. Leaf Node
3. Internal Node
4. Jump Node

Decision Trees

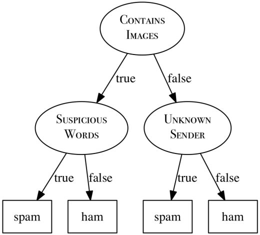
Suppose you have an email dataset with three Boolean attributes - Suspicious Words, Uknown Sender, and Contains Images, and an Boolean output which could be spam or ham You create the decision tree below.



What would be the predicted classification for Suspicious Words = true, Unknown Sender = True, and Contains Images = False

Decision Trees

Suppose you have an email dataset with three Boolean attributes - Suspicious Words, Uknown Sender, and Contains Images, and an Boolean output which could be spam or ham You create the decision tree below.



What would be the predicted classification for Suspicious Words = true, Unknown Sender = True, and Contains Images = False

Spam

Entropy

What is the definition of Entropy of a dataset containing k classes, $1, 2, \ldots, k$, where each class i has probability p_i.

Entropy

What is the definition of Entropy of a dataset containing k classes, $1, 2, \ldots, k$, where each class i has probability p_i.

$$H(\text{dataset}) = -\sum_{i=1}^{k} p_i log_2(p_i)$$

Entropy

What is the entropy of a set of 52 playing cards if we only distinguish between the cards based on their suit $\{\heartsuit, \clubsuit, \diamondsuit, \spadesuit\}$?

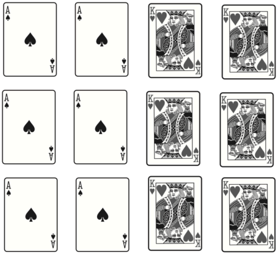
Entropy

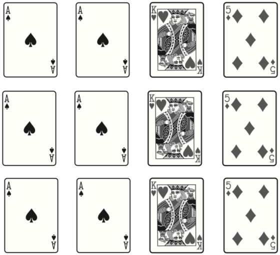
What is the entropy of a set of 52 playing cards if we only distinguish between the cards based on their suit $\{\heartsuit, \clubsuit, \diamondsuit, \spadesuit\}$?

$$
\begin{aligned}
H(suit) &= -\sum_{l \in \{\heartsuit,\clubsuit,\diamondsuit,\spadesuit\}} P(suit = l) \times log_2(P(suit = l)) \\
&= -\big((P(\heartsuit) \times log_2(P(\heartsuit))) + (P(\clubsuit) \times log_2(P(\clubsuit))) \\
&\quad + (P(\diamondsuit) \times log_2(P(\diamondsuit))) + (P(\spadesuit) \times log_2(P(\spadesuit))) \big) \\
&= -\Big(\left({}^{13}/_{52} \times log_2({}^{13}/_{52})\right) + \left({}^{13}/_{52} \times log_2({}^{13}/_{52})\right) \\
&\quad + \left({}^{13}/_{52} \times log_2({}^{13}/_{52})\right) + \left({}^{13}/_{52} \times log_2({}^{13}/_{52})\right) \Big) \\
&= -\big((0.25 \times -2) + (0.25 \times -2) \\
&\quad + (0.25 \times -2) + (0.25 \times -2) \big) \\
&= 2 \; bits
\end{aligned}
$$

Entropy

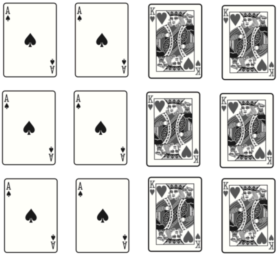
Which of the following has lower entropy?

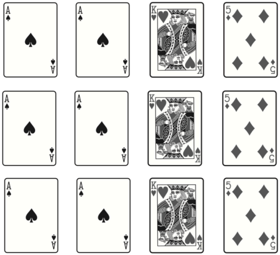




Entropy

Which of the following has lower entropy?





Left one has entropy of 1.00 and right has entropy of 1.50

Information Gain

In the following dataset, which attribute out of X, Y, and Z would give the best information gain

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II

Information Gain

In the following dataset, which attribute out of X, Y, and Z would give the best information gain

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II

Best Attribute: Y

Information Gain

- Go through some examples of creating decision trees using ID3

Information Gain

- Go through some examples of creating decision trees using ID3
- No need to study alternate information measures, noise, overfitting, etc

Outline

Probability

Conditional Probability:

$$P(X = a|Y = b) = \frac{P(X = a, Y = b)}{P(Y = b)}$$
$$= \frac{P(X = a)P(Y = b|X = a)}{P(Y = b)}$$

Probability

10% of patients entering a doctor's office have liver disease, and 5% of patients walking in are alcoholics. Of those that are diagnosed with liver disease, 7% are alcoholics. Given a patient walking into the office is alcoholic, what is the probability that he will be diagnosed with liver disease.

Probability

10% of patients entering a doctor's office have liver disease, and 5% of patients walking in are alcoholics. Of those that are diagnosed with liver disease, 7% are alcoholics. Given a patient walking into the office is alcoholic, what is the probability that he will be diagnosed with liver disease.

A = having liver disease
B = being alcoholic

$$
\begin{aligned}
P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\
&= \frac{0.07 \times 0.10}{0.05} \\
&= 0.14
\end{aligned}
$$

14%

Probability

From a standard deck of playing cards, someone draws a random card and tells you it's a face card. What is the probability that it is a king of any suit?
Hint: J, Q, K are called face cards.

Probability

From a standard deck of playing cards, someone draws a random card and tells you it's a face card. What is the probability that it is a king of any suit?
Hint: J, Q, K are called face cards.

A = card being a king
B = drawing a face card

$$
\begin{aligned}
P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\
&= \frac{1.00 \times (4/52)}{(12/52)} \\
&= 1/3
\end{aligned}
$$

33.3%

Probability

In an experiment if there are n possible outcomes which are denoted by random variable X taking values $X_1, X_2, \ldots, X_n$ with respected probability values $p_1, p_2, \ldots, p_n$, the expected value of X is :

$$E(X) = \mu(X) = \sum_{i=1}^{n} p_i X_i$$

and variance of X is:

$$\begin{aligned} var(X) &= E[(X - \mu)^2] \\ &= E[X^2] - E[X]^2 \end{aligned}$$

Expected Value

An insurance company sells a policy that costs \$150, and pays \$5000 for major accident, and \$1000 for a minor accident only once. The probability of major accident is 0.005 and that of minor incident is 0.08. What is the expected value of the profit/loss for each policy sold by the company.

Expected Value

An insurance company sells a policy that costs \$150, and pays \$5000 for major accident, and \$1000 for a minor accident only once. The probability of major accident is 0.005 and that of minor incident is 0.08. What is the expected value of the profit/loss for each policy sold by the company.

There could be three cases for the company

1. No accident, profit = \$150, probability $= 1 - 0.005 - 0.08 = 0.915$
2. Major accident, profit = \$ (150 - 5000) = -\$4850, probability = 0.005
3. Minor accident, profit = \$ (150 - 1000) = -\$850, probability = 0.08

$$
\begin{aligned}
E(X) &= 150 * 0.915 - 4850 * 0.005 - 850 * 0.08 \\
&= 45
\end{aligned}
$$

Variance

In a game, the probability of success is 0.6, and for a win you are paid \$1 and for a loss you lose \$1. What is the variance of outcome?

Variance

In a game, the probability of success is 0.6, and for a win you are paid \$1 and for a loss you lose \$1. What is the variance of outcome?

$E(X) = p = 0.6$

X	+1	-1
p(X)	0.6	0.4
$(X - E(X))^2)$	$(1 - 0.6)^2$	$(-1 - 0.6)^2$

$$\begin{aligned} \text{var}(X) &= 0.6 * (0.4)^2 + 0.4 * (1.6)^2 \\ &= 1.12 \end{aligned}$$

Outline

Naive Bayes Classifier

Suppose there are two classed C_1 and C_2, and we are given data $X = (X_1, X_2, \ldots, X_n)$, the probability of it belonging to C_1 and C_2 is:

Naive Bayes Classifier

Suppose there are two classed C_1 and C_2, and we are given data $X = (X_1, X_2, \ldots, X_n)$, the probability of it belonging to C_1 and C_2 is:

$$P(C_1|X) = \frac{P(X|C_1)P(C_1)}{P(X)} \propto P(X|C_1)P(C_1)$$

$$P(C_2|X) = \frac{P(X|C_2)P(C_2)}{P(X)} \propto P(X|C_2)P(C_2)$$

Naive Bayes Classifier

Suppose there are two classed C_1 and C_2, and we are given data $X = (X_1, X_2, \ldots, X_n)$, the probability of it belonging to C_1 and C_2 is:

$$P(C_1|X) = \frac{P(X|C_1)P(C_1)}{P(X)} \propto P(X|C_1)P(C_1)$$

$$P(C_2|X) = \frac{P(X|C_2)P(C_2)}{P(X)} \propto P(X|C_2)P(C_2)$$

The Maximum a Posteriori (MAP) hypothesis compares $P(X|C_i)P(C_i)$ for each class C_i and assigns X to the class that has the maximum value.

Naive Assumption

In many cases, it's not possible to evaluate $P(X|C_1) = P((X_1, X_2, \ldots, X_n)|C_i)$ because of lack of training data.

In such cases, we make the assumption that each of the features $X_1, X_2, \ldots, X_n$ is independent of each other for every class C_i (called the naive Bayes assumption)

$$\begin{aligned} P(X|C_1) &= P((X_1, X_2, \ldots, X_n)|C_i) \\ &= P(X_1|C_i)P(X_2|C_i)\ldots P(X_n|C_i) \end{aligned}$$

Naive Bayes Classifier

We have a small dataset consisting of three classes *well, cold, allergy* and three words *sneeze, cough, fever*. The distribution of the words for each class is shown below:

Prob	**Well**	**Cold**	**Allergy**
$P(c_i)$	0.9	0.05	0.05
$P(\text{sneeze}\|c_i)$	0.1	0.9	0.9
$P(\text{cough}\|c_i)$	0.1	0.8	0.7
$P(\text{fever}\|c_i)$	0.01	0.7	0.4

What would be the predicted tag for the a text containing words W $=$ *sneeze, cough,* $\neg$ *fever*

Naive Bayes Classifier

$$\begin{aligned} P(well|W) &= P(sneeze|well)P(cough|well)P(\neg fever|well)P(well) \\ &= 0.1 \times 0.1 \times 0.99 \times 0.9 \\ &= 0.0089 \end{aligned}$$

$$\begin{aligned} P(cold|W) &= P(sneeze|cold)P(cough|cold)P(\neg fever|cold)P(cold) \\ &= 0.9 \times 0.8 \times 0.3 \times 0.05 \\ &= 0.01 \end{aligned}$$

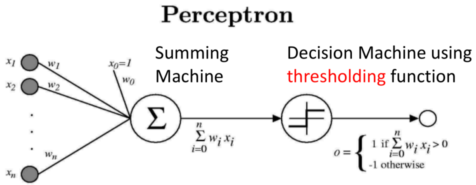
$$\begin{aligned} P(allergy|W) &= P(sneeze|allergy)P(cough|allergy)P(\neg fever|allergy)P(allergy) \\ &= 0.9 \times 0.7 \times 0.6 \times 0.05 \\ &= 0.019 \end{aligned}$$

The predicted class is **allergy**

Outline

Perceptron

Perceptron is a single unit of computation, that consists of an addition operation followed by an activation operation.



$$o(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } w_0 + w_1 x_1 + \cdots + w_n x_n > 0 \\ -1 & \text{otherwise.} \end{cases}$$

Sometimes we'll use simpler vector notation:

$$o(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \cdot \vec{x} > 0 \\ -1 & \text{otherwise.} \end{cases}$$

Note the classification rule above.

Perceptron

Another way of binary classification

Linear Classification

- Two equivalent ways:

$$w^T x > \theta \text{ where } w = \begin{pmatrix} w_1 \\ w_2 \\ \dots \\ w_n \end{pmatrix} x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \text{ for class +1}$$

$$w^T x > 0 \text{ where } w = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \dots \\ w_n \end{pmatrix} x = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \text{ for class +1}$$

Note the classification rule above.

Perceptron

Go through **AND, OR** classification using Perceptron from class slides.
You could be asked similar questions for AND NOT, NOT OR, etc

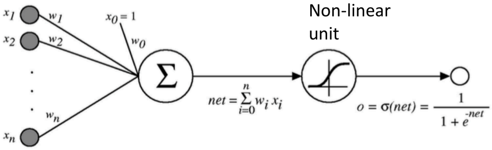
Perceptron

Go through **AND, OR** classification using Perceptron from class slides.
You could be asked similar questions for AND NOT, NOT OR, etc

Outline

Neural Networks

In neural nets, each neuron uses a non-linear activation function, such as sigmoid.



$\sigma(x)$ is the sigmoid function

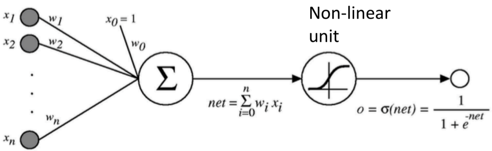
$$\frac{1}{1+e^{-x}}$$

Nice property: $\frac{d\sigma(x)}{dx} = \sigma(x)(1-\sigma(x))$ Really useful result

Note the properties of sigmoid and its derivative.

Neural Networks

In neural nets, each neuron uses a non-linear activation function, such as sigmoid.



$\sigma(x)$ is the sigmoid function

$$\frac{1}{1+e^{-x}}$$

Nice property: $\frac{d\sigma(x)}{dx} = \sigma(x)(1-\sigma(x))$ Really useful result

Note the properties of sigmoid and its derivative.

Backpropagation Algorithm for Sigmoid Activation

Backpropagation Algorithm

Initialize all weights to small random numbers

Until convergence, Do

For each training example, Do

1. Input it to network and compute network outputs — Forward pass
2. For each output unit k

$$\delta_k \leftarrow o_k(1 - o_k)(t_k - o_k)$$

3. For each hidden unit h

$$\delta_h \leftarrow o_h(1 - o_h) \sum_{k \in outputs} w_{h,k}\delta_k$$

4. Update each network weight $w_{i,j}$ — Backward pass

$$w_{i,j} \leftarrow w_{i,j} + \Delta w_{i,j}$$

where $\Delta w_{i,j} = \eta \delta_j x_{i,j}$

Some possible questions

Go through a simple example of backpropagation from class slides and questions.

Calculate square error at the output layer

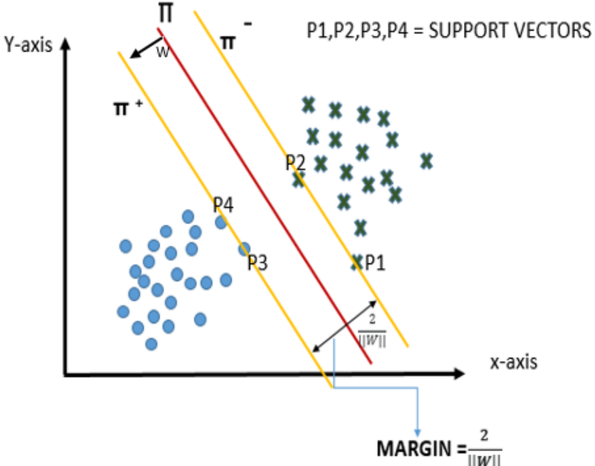
Compute weight updates for backward pass

Given number of neurons in each layer, find number of connections.

Outline

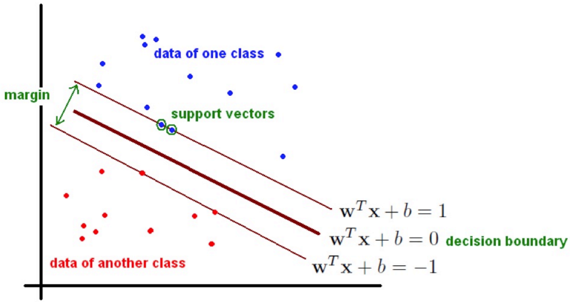
Support Vector Machines

Aim of SVM is to find the maximum margin classifier.



Support Vector Machines

SVM line is the center of two parallel lines passing through closest points on either sides, known as **support vectors**.



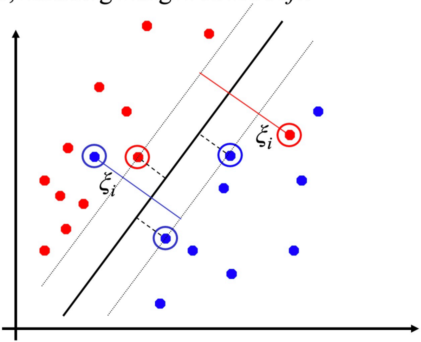
Source: https://web.mit.edu/zoya/www/SVM.pdf

Support Vector Machines

Non-separable case

What if the training set is not linearly separable?

Slack variables ξ_i can be added to allow misclassification of difficult or noisy examples, resulting margin called *soft*.



Source: Based on slides from Prof. Ray Mooney

Support Vector Machines

Learn to compute following:

- Distance between a point and a line (or hyperplane) in scalar and vector formats.
- Distance between two parallel lines (or hyperplanes) in scalar and vector formats.
- Distance between two points in scalar and vector formats.
- Setting up of SVM optimization equation.
- Kernel Method
- Non-separable case, and C parameter