survival8

BITS WILP Machine Learning Quiz-1 2017-H2

Index Subjects Mail Us

BITS WILP Machine Learning (ISZC464) Quiz-1 2017-H2

Q1.
Which of the following statements are true in context of Graphical models?
Select one or more:
a. None of the above.
b. Bayesian Belief networks describe conditional independence among subsets of variables.
c. Bayes network represents the joint probability distribution over a collection of random variable.

d. Each node denotes a random variable.

Answer: (B, C, D)
A **Bayesian network**, **Bayes network**, **belief network**, **Bayes(ian) model** or **probabilistic directed acyclic graphical model** is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG).

> A Bayesian belief network describes the probability distribution governing a set of variables by specifying a set of conditional independence assumptions along with a set of conditional probabilities. In contrast to the naive Bayes classifier, which assumes that *all* the variables are conditionally independent given the value of the target variable, Bayesian belief networks allow stating conditional independence assumptions that apply to *subsets* of the variables. Thus, Bayesian belief networks provide an intermediate approach that is less constraining than the global assumption of conditional independence made by the naive Bayes classifier, but more tractable than avoiding conditional independence assumptions altogether. Bayesian belief networks are an active focus of current research, and a variety of algorithms have been proposed for learning them and for using them for inference.

Q2.
Assuming log base 2, the entropy of a binary feature with p(x=1) = 0.5 is
Select one:
a. 0.75
b. 0
c. 0.25
d. 1
e. 0.5
Answer: (D) It is '1'.

Pages

- Postings Index
- Index of BITS WILP Exam Papers and Content
- Index of Lessons in Technology
- Index of Guest Interviews
- Downloads
- Book Requests

Blog Archive

▼ 2020 (31)
▼ May (1)
Covid-19 and response of IT companies (by Divjot S...
► April (6)
► March (12)
► February (6)
► January (6)
► 2019 (48)
► 2018 (31)
► 2017 (15)
► 2016 (6)

Popular Posts

You Are a Badass. How to stop doubting your greatness and start living an awesome life (Jen Sincero, 2013)
INTRODUCTION The language used in the book extremely funny and Jen Sincero still makes sure that she m...

Covid-19 and response of IT companies (by Divjot Singh)
As the Covid-19 pandemic ravages the world, many domains like airlines, tourism and services...

Innovation to beat the Coronavirus (Covid19)
Coronovirus' Exponential growth and decline In the first phase of the pandemic, we saw a...

Download fiction books (March 2018)
Download fiction books for free: Link for Google Dr...

Life Lessons By Steve Jobs

Entropy

Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x)$

$$p(x) = Pr\{X = x\}, \quad x \in \mathcal{X}$$

The *entropy* of the variable X is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

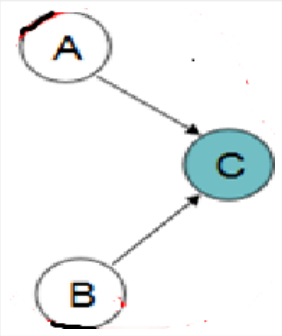
The logarithm can be in any base, but normally base 2 is used. The unit of the entropy is then called *bits*. If we use base b in the logarithm, we denote the entropy by $H_b(X)$. We can easily convert between entropies in different bases

$$H_b(X) = \log_b a \cdot H_a(X)$$

By convention $0 \log 0 = 0$, since $y \log y \to 0$ as $y \to 0$.

The entropy is a measure of the information content of a random variable.

Q3.
Which of the following statement are true for the given graphical model.



Select one or more:
a. A is conditionally independence of B given C.
b. B is conditionally independence of A given C.
c. B is not conditionally independence of A given C.
d. A is not conditionally independence of B given C.
Answer: C, D
(B is not conditionally independence of A given C., A is not conditionally independence of B given C.)

Q4.
Let X be random variable and let Y=aX+b, where a and b are given scalars. Then which of the following statements are true. (E[Z] states the expected value of Z)
Select one or more:
a. E[Y]=(a/b)*E[X]
b. E[Y] = E[X]
c. E[Y]=a*E[X]+b
d. E[Y]=a*b*E[X]
Answer: (C)

Expected value of a constant is constant [edit]

If c is a constant random variable, then $\mathrm{E}[c] = c$. This implies that for any random variable X, $\mathrm{E}[\mathrm{E}[X]] = \mathrm{E}[X]$

Linearity [edit]

The expected value operator (or **expectation operator**) $\mathrm{E}[\cdot]$ is linear in the sense that

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y],$$

$$\mathrm{E}[aX] = a\,\mathrm{E}[X],$$

where X and Y are (arbitrary) random variables, and a is a scalar.

If a and b are constants then $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$

$\mathrm{E}(aX+b) = a\,\mathrm{E}(X) + b$
$\mathrm{Var}(aX+b) = \mathrm{E}[(aX+b-(a\,\mathrm{E}(X)+b))^2] = \mathrm{E}(a^2(X-\mathrm{E}(X))^2) = a^2\mathrm{E}((X-\mathrm{E}(X))^2) = a^2\mathrm{Var}(X)$

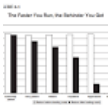
The square root of Var(X) is called the standard deviation of X.
SD(X) = sqrt(Var(X)): measures scale of X.

Q5.
When we can use Expectation maximization algorithm.

Steve Jobs' last words will change your views on life. The billionaire passed away at the ...

Effects of news and world events on Nifty50 and stock market
Day: 10th Aug 2017 Sensex tanks 267 points. Nifty hits one-month low. 1. Market outlook: ...

Why Bill Gates would raise chickens
I'm excited about the poverty-fighting power of poultry. If you were living on $2 a day, wh...



Intelligent investor (Ben Graham & Jason Zweig, 4e)
Reading from "A Note About Benjamin Graham by Jason Zweig" Here are Graham...

The Essays Of Warren Buffett (Lessons For Corporate America)
INTRODUCTION Buffett has applied the traditional principles as chief executive officer of Berkshi...

How To Talk TO Anyone (92 Little Tricks For Big Success In Relationships, by Leil Lowndes) - Book Summary
There are two kinds of people in this life: Those who walk into a room and say, "Well, here I...

About Me

Ashish Jain

View my complete profile

Select one or more:
a. None of the these.
b. Unsupervised clustering (target value unobservable).
c. Data is only partially observable.
d. Supervised Learning (some instance attributes unobservable).
Answer: B, C, D

Expectation Maximization (EM) Algorithm

- When to use
  - ✓ Data is only partially observable
  - ✓ Unsupervised clustering (target value unobservable)
  - ✓ Supervised Learning (some instance attributes unobservable)
- Some uses
  - ✓ Train Bayesian Belief Networks
  - ✓ Unsupervised clustering
  - ✓ Learning Hidden Markov Models

Q6.
Which of the following statements are true?
Select one or more:
a. Maximum a Posteriori estimation seek the estimate of Θ that is most probable, given the observed data, plus background assumptions about its value.
b. Maximum Likelihood estimation seek the estimate of Θ that is most probable, given the observed data, plus background assumptions about its value.
c. Maximum Likelihood estimation seek an estimate of Θ that maximizes the probability of the observed data.
d. Maximum a Posteriori estimation seek an estimate of Θ that maximizes the probability of the observed data.
Answer: A, C

MAP principle: We should choose the value of θ that is most probable, given the observed data D and our prior assumptions summarized by P(θ); that is

$$\hat{\theta}^{MAP} = \arg\max_{\theta} P(\theta|D)$$

...

Maximum Likelihood Estimation (MLE)

- We should choose the value of θ that makes data set, D most probable

$$\hat{\theta}^{MLE} = \arg\max_{\theta} P(D|\theta)$$

In MLE, we don't have prior knowledge, as in the example of a toss of coin, about the coin whether it is biased or unbiased. We arrive at Theta based on the data. While in MAP, we incorporate our prior knowledge:

Second Algorithm allow us to incorporate our **prior knowledge** about coins by adding any number of imaginary coin flips resulting in heads and tails.

Let assume, γ_1 denotes imaginary heads
γ_0 denotes imaginary tails

Considering this prior knowledge, now the Θ̈ can be estimated as follows:

$$\ddot{\Theta} = \alpha_1 + \gamma_1 / (\alpha_1 + \gamma_1 + \alpha_0 + \gamma_0)$$

Q7.
If X is a vector of n attributes and Y is boolean valued label. How many different functions are possible? (2^n represents 2^n)
a. 2^{2^n}
b. 2^{n^2}
c. 2^n
d. 2n
Answer: A

It is 2^(2^n)
If X has two attributes x1 and x2, then # of observations one has to take are (x1=0, x2=0, y), (x1=0, x2=1, y), (x1=1, x2=0, y), (x1=1, x2=1, y). 'n' attributes means 2^n states.

Number of functions would be: 2^(n^2)

X1, X2, Y
Function: 1
0, 0, 0
0, 1, 0
1, 0, 0
1, 1, 0
Function: 2
0, 0, 0
0, 1, 0
1, 0, 0
1, 1, 1

Run 1:
Input: X1, X2
0, 0
0, 1
1, 0
1, 1
Output: 0,0,0,0

Run 2:
Input: X1, X2
0, 0
0, 1
1, 0
1, 1
Output: 0,0,0,1

Run 3:
Input: X1, X2
0, 0
0, 1
1, 0
1, 1
Output: 0,0,1,0

Run 4:
Input: X1, X2
0, 0
0, 1
1, 0
1, 1
Output: 0,0,1,1

Run 5: Output: 0,1,0,0. Run 6: Output: 0,1,0,1. Run 7: Output: 0,1,1,0. Run 8: Output: 0,1,1,1
Run 9: Output: 1,0,0,0. Run 10: Output: 1,0,0,1. Run 11: Output: 1,0,1,0. Run 12: Output: 1,0,1,1
Run 13: Output: 1,1,0,0. Run 14: Output: 1,1,0,1. Run 15: Output: 1,1,1,0. Run 16: Output: 1,1,1,1

This can be understood as there will be 2^n rows in one truth table for X1, X2. Now, assume Y to be a vector of length 2^n, number of states it can take = 2^(2^n).

Q8.
Which of the following statements are true?
Select one or more:
a. To infer posterior probability, Bayesian linear regression uses Naïve Bayes principle.
b. None of these
c. Bayesian linear regression cannot be used for classification.
d. In Bayesian linear regression Prior can be used for regularization.
Answer: A, D
In Bayesian linear regression Prior can be used for regularization., To infer posterior probability, Bayesian linear regression uses Naïve Bayes principle.

Bayesian Linear Regression

innovate achieve

- $D = \{(x_1,y_1), (x_2,y_2),\ldots,(x_N,y_N)\}$ where $x_i \in \mathbb{R}^D$ and $y \in \mathbb{R}$
- Model : $Y_1, Y_2, \ldots, Y_N$ independent given $\mathbf{w}$, $Y \sim \mathcal{N}(w^T x_i, \beta)$ [β is precision; $\beta = 1/\sigma^2$]
- $\mathbf{w} \sim \mathcal{N}(0, \alpha^{-1}I)$ where $\mathbf{w} = (w_0, w_1, \ldots, w_D)^T$
- Assume β and α are known
  - ✓ therefore only unknown parameter is $\mathbf{w}$
- **Likelihood:**

$$p(\boldsymbol{D}|\boldsymbol{w}) \alpha\ exp(-\frac{\beta}{2}(\boldsymbol{y}-\boldsymbol{Qw})^T(\boldsymbol{y}-\boldsymbol{Qw}))$$

- **Posterior:**

$$p(\boldsymbol{w}|\boldsymbol{D}) \alpha\ p(\boldsymbol{D}|\boldsymbol{w})\ p(\boldsymbol{w})$$

...

Posterior of w

innovate achieve lead

$$p(\boldsymbol{w}|\boldsymbol{D}) \alpha\ p(\boldsymbol{D}|\boldsymbol{w})\ p(\boldsymbol{w})$$

$$p(\boldsymbol{w}|\boldsymbol{D}) \alpha\ exp\left(-\frac{\beta}{2}(\boldsymbol{y}-\boldsymbol{Qw})^T(\boldsymbol{y}-\boldsymbol{Qw})\right)\ exp(-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w})$$

$$p(\boldsymbol{w}|\boldsymbol{D}) \alpha\ exp\left(-\frac{\beta}{2}(\boldsymbol{y}-\boldsymbol{Qw})^T(\boldsymbol{y}-\boldsymbol{Qw}) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\right)$$

$$p(\boldsymbol{w}|\boldsymbol{D}) \alpha\ exp\left(-\frac{1}{2}(\beta(\boldsymbol{y}-\boldsymbol{Qw})^T(\boldsymbol{y}-\boldsymbol{Qw}) + \alpha\boldsymbol{w}^T\boldsymbol{w})\right)$$

$$p(\boldsymbol{w}|\boldsymbol{D}) \alpha\ exp\left(-\frac{1}{2}\left(\beta(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T\boldsymbol{Q}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{Q}^T\boldsymbol{Qw})\right) + \alpha\boldsymbol{w}^T\boldsymbol{w})\right)\ [-2\boldsymbol{w}^T\boldsymbol{Q}^T\boldsymbol{y} = -\boldsymbol{y}^T\boldsymbol{Qw} - (\boldsymbol{Qw})^T\boldsymbol{y}\ ]$$

$$p(\boldsymbol{w}|\boldsymbol{D}) \alpha\ exp\left(-\frac{1}{2}\left((\beta\boldsymbol{y}^T\boldsymbol{y} - 2\beta\boldsymbol{w}^T\boldsymbol{Q}^T\boldsymbol{y} + \beta\boldsymbol{w}^T\boldsymbol{Q}^T\boldsymbol{Qw})\right) + \alpha\boldsymbol{w}^T\boldsymbol{w})\right)$$

$$p(\boldsymbol{w}|\boldsymbol{D}) \alpha\ exp\left(-\frac{1}{2}((\beta\boldsymbol{y}^T\boldsymbol{y} - 2\beta\boldsymbol{w}^T\boldsymbol{Q}^T\boldsymbol{y} + \boldsymbol{w}^T(\beta\boldsymbol{Q}^T\boldsymbol{Q} + \alpha\boldsymbol{I})\boldsymbol{w}\right)$$

...

Posterior of w

$$p(\boldsymbol{w}|\boldsymbol{D}) \alpha\ exp\left(-\frac{1}{2}((\beta\boldsymbol{y}^T\boldsymbol{y} - 2\beta\boldsymbol{w}^T\boldsymbol{Q}^T\boldsymbol{y} + \boldsymbol{w}^T(\beta\boldsymbol{Q}^T\boldsymbol{Q} + \alpha\boldsymbol{I})\boldsymbol{w}\right) \quad :::1$$

Below we are writing a multi-variate Gaussian distribution:

Completing the square:

$$\mathcal{N}(\boldsymbol{\mu}, \Lambda^{-1}) \alpha\ \boldsymbol{exp}(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})^T\Lambda(\boldsymbol{w}-\boldsymbol{\mu}))$$

$$\mathcal{N}(\boldsymbol{\mu}, \Lambda^{-1}) \alpha\ \boldsymbol{exp}(-\frac{1}{2}(\boldsymbol{w}^T\Lambda\boldsymbol{w} - \boldsymbol{w}^T\Lambda\boldsymbol{\mu} - \boldsymbol{\mu}^T\Lambda\boldsymbol{w} + \boldsymbol{\mu}^T\Lambda\boldsymbol{\mu}))$$

$$\mathcal{N}(\boldsymbol{\mu}, \Lambda^{-1}) \alpha\ \boldsymbol{exp}(-\frac{1}{2}(\boldsymbol{w}^T\Lambda\boldsymbol{w} - 2\boldsymbol{w}^T\Lambda\boldsymbol{\mu} + \boldsymbol{\mu}^T\Lambda\boldsymbol{\mu})) \quad :::2$$

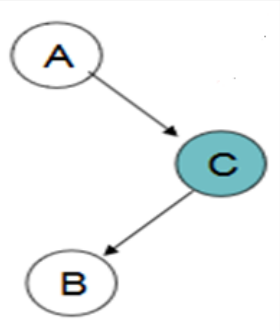
On matching expression (1) and (2), we get:

$$\Lambda = \beta\boldsymbol{Q}^T\boldsymbol{Q} + \alpha\boldsymbol{I}$$

$$\mu = \beta\Lambda^{-1}\boldsymbol{Q}^T\boldsymbol{y} \quad \text{(Using } \boldsymbol{w}^T\Lambda\mu = \beta\boldsymbol{w}^T\boldsymbol{Q}^T\boldsymbol{y})$$

These slides show the derivation of posterior probability using Bayes theorem, and here all probabilities are represented by multivariate Gauss distribution.

Q9.
Which of the following statement are true for the given graphical model?



Select one or more:
a. B is not conditionally independence of A given C.
b. A is conditionally independence of B given C.
c. B is conditionally independence of A given C.
d. A is not conditionally independence of B given C.
Answer: B, C

Conditional independence

- Let A, B, and C be events. A and B are *conditionally independent given* C iff

$$P(A \mid C) = P(A \mid B \cap C)$$

or, equivalently, iff

$$P(A \cap B \mid C) = P(A \mid C) \times P(B \mid C)$$

- If A and B are conditionally independent, then once we learn C, learning B gives us no *additional* information about A.
- Two random variables X and Y are conditionally independent given Z if for all x, y, and z

$$p_{X|Z}(x, z) = p_{X|YZ}(x, y, z)$$

In this case we write $X \perp\!\!\!\perp Y \mid Z$. This also corresponds to (perhaps infinitely) many event conditional independencies.

...

Head to Tail

- Prove A conditional independence of B given C?

Proved in : P(A,B | C) = P(A,B,C)/P(C) = P(A)*P(C|A)*P(B|A)/P(C) = P(A|C) * P(B|C)
Hence, P(A,B | C) = P(A|C) * P(B|C)

This implies "A and B are conditionally independent given C".

Q10.
Smoothing can be used in which of the following cases:
Select one or more:
a. When likelihood estimates zero probability
b. When test error and training error are very different
c. None of the above
d. When learning algorithm result in very rough function
Answer: A
(When likelihood estimates zero probability.)
Could someone explain Laplacian smoothing (or 1-up smoothing)?
Ans: Suppose you are looking at outcomes of a die. Let us say you get the following outcomes of each number, in 10 throws:
One : 1
Two : 3
Three : 1
Four : 0
Five : 3
Six : 2

Now, the probabilities without the smoothing are
One : 1/10
Two : 3/10
Three : 1/10
Four : 0/10
Five : 3/10
Six : 2/10

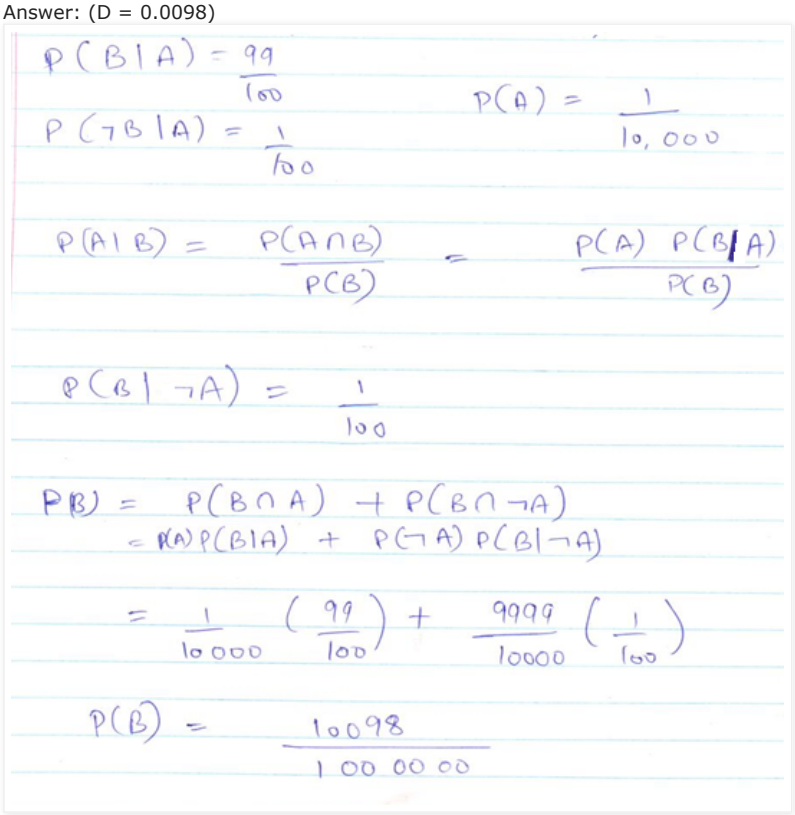
The sums of probabilities is (of course) 1.
To smoothen out, we add '1' to numerators. Now we need to add "something" to the denominator such that the sum remains 1.
So,
(1+1+3+1+1+1+0+1+3+1+2+1) / (10+*K*) = 1

This gives K=6. Now note that if you had zero throws, the probabilities are all 1/6. These are called the "prior probabilities" - our prior assumption of the outcomes. We initially believe all of them are equally likely.
And K=6 is essentially the no. of classes!
(URL: https://www.quora.com/Could-someone-explain-Laplacian-smoothing-or-1-up-smoothing)

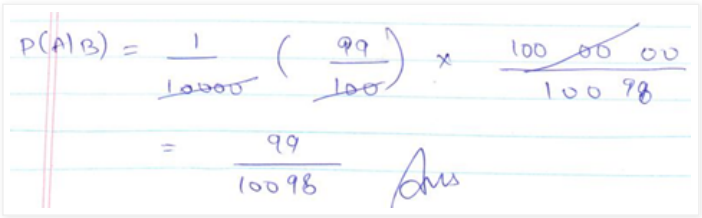
Q11.
Which of the following statements are true in context of decision trees?
Select one or more:
a. Capable in classifying non-linearly separable data.
b. None of these.
c. Capable in classifying linearly separable data.
d. It is always possible to get zero training error.
Answer: A, C, D
"Zero training error": means decision tree can give a model that will give correct output for all the traning data.
An attribute is a discreete valued variable. While traversing a decision tree downwards based on attributes, it is always possible to arrive at a label (decision (yes, no)).

Q12.
Let a probability of disease is 1 in 10,000 and the test accuracy of the disease is 99 %. Let event A is the event you have this disease, and event B is the event that you test positive. Given test is positive what is the probability that disease is actually present? Precisely you need calculate probability P(A|B)
Select one:
a. 0.0990
b. 0.0988
c. 0.9902

d. 0.0098
Answer: (D = 0.0098)



...



Q13.
In context of Bais-Variance decomposition which of the following statements are true?
Select one or more:
a. High bais implies high vairiance in the out of sample error.
b. High variance implies less bias in the out of sample error.
c. Less bais implies less variance in the out of sample error.
d. Bais-Variance analysis help us to quantify out of sample error.
Answer: B, D
(From L6-Part2 last slide)
(URL: http://www.stat.cmu.edu/~ryantibs/advmethods/notes/errval.pdf)

~~fit on this training set.~~ We'll look at the expected test error, conditional on $X = x$ for some arbitrary input x,

$$\begin{aligned}\mathbb{E}[\text{TestErr}(\hat{r}(x))] &= \mathbb{E}\big[\big(Y - \hat{r}(x)\big)^2 \big| X = x\big] \\ &= \mathbb{E}\big[\big(Y - r(x)\big)^2 \big| X = x\big] + \mathbb{E}\big[\big(r(x) - \hat{r}(x)\big)^2 \big| X = x\big] \\ &= \sigma^2 + \mathbb{E}\big[\big(r(x) - \hat{r}(x)\big)^2\big].\end{aligned}$$

The first term is just a constant, σ^2, and is the *irreducible error* (sometimes referred to as the *Bayes error*). The second term can be further decomposed as

$$\begin{aligned}\mathbb{E}\big[\big(r(x) - \hat{r}(x)\big)^2\big] &= \big(\mathbb{E}[\hat{r}(x)] - r(x)\big)^2 + \mathbb{E}\big[\big(\hat{r}(x) - \mathbb{E}[\hat{r}(x)]\big)^2\big] \\ &= \text{Bias}(\hat{r}(x))^2 + \text{Var}(\hat{r}(x)),\end{aligned}$$

the first term being the squared *estimation bias* or simply *bias*, $\text{Bias}(\hat{r}(x)) = \mathbb{E}[\hat{r}(x)] - r(x)$, and the second term being the *estimation variance* or simply *variance*. Therefore, altogether,

$$\mathbb{E}[\text{TestErr}(\hat{r}(x))] = \sigma^2 + \text{Bias}(\hat{r}(x))^2 + \text{Var}(\hat{r}(x)), \tag{2}$$

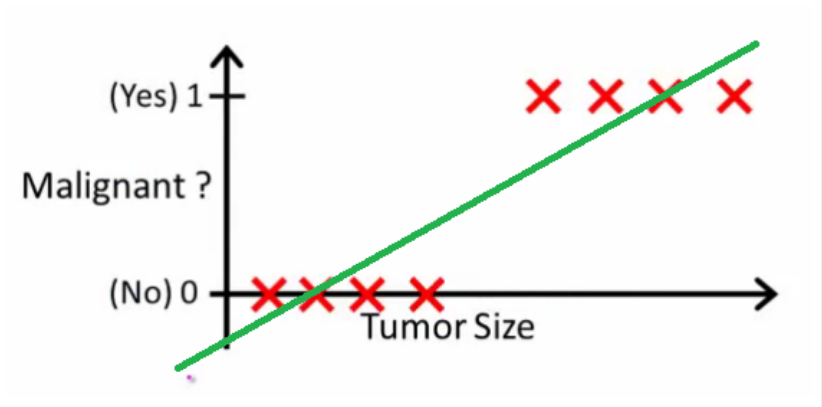
which is called the *bias-variance decomposition* or *bias-variance tradeoff*

- From the bias-variance tradeoff (2), we can see that even if our prediction is unbiased, i.e., $\mathbb{E}[\hat{r}(x)] = r(x)$, we can still incur a large error if it is highly variable. Meanwhile, even when our prediction is stable and not variable, we can incur a large error if it is badly biased
- There is a tradeoff here, but it need it is really never be one-to-one; i.e., in some cases, it can be worth sacrificying a little bit of bias to gain large decrease in variance, and in other cases, vice versa
- Typical trend: underfitting means high bias and low variance, overfitting means low bias but high variance. E.g., think about k in k-nearest-neighbors regression: relatively speaking, how do the bias and variance behave for small k, and for large k?

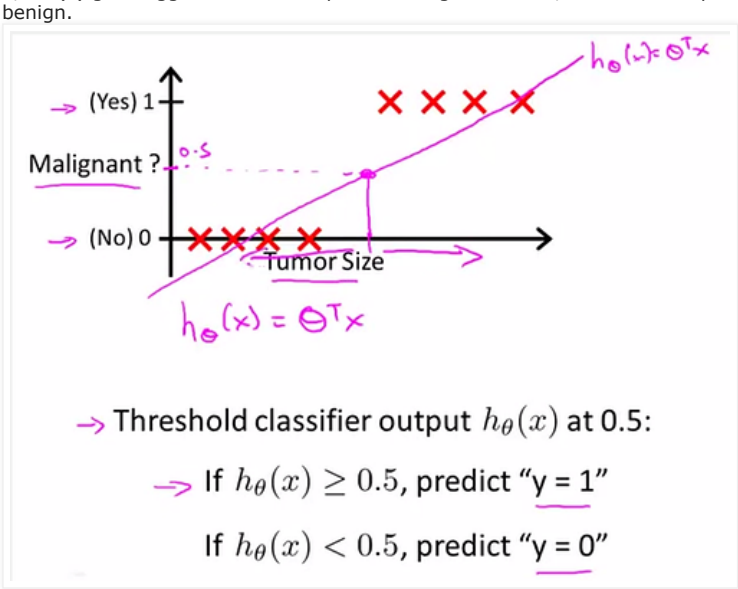
Q14.
In context of linear regression, which of the following statements are true?
Select one or more:
a. You can use linear regression for classification.
b. It is not possible to get zero training error, if there are few samples used in training.
c. It is not possible to get zero test error, if there are few samples used in training.
d. You cannot use linear regression for classification.
Answer: A, C
Training error is the error that you get when you run the trained model back on the training data. Remember that this data has already been used to train the model and this necessarily doesn't mean that the model once trained will accurately perform when applied back on the training data itself.

Test error is the error when you get when you run the trained model on a set of data that it has previously never been exposed to. This data is often used to measure the accuracy of the model before it is shipped to production.
.....
URL: https://stats.stackexchange.com/questions/22381/why-not-approach-classification-through-regression
QUESTION: Some material I've seen on machine learning said that it's a bad idea to approach a classification problem through regression. But I think it's always possible to do a continuous regression to fit the data and truncate the continuous prediction to yield discrete classifications. So why is it a bad idea?

ANSWER:
"..approach classification problem through regression.." by "regression" I will assume you mean linear regression, and I will compare this approach to the "classification" approach of fitting a logistic regression model.
Before we do this, it is important to clarify the distinction between regression and classification models. Regression models predict a continuous variable, such as rainfall amount or sunlight intensity. They can also predict probabilities, such as the probability that an image contains a cat. A probability-predicting regression model can be used as part of a classifier by imposing a decision rule - for example, if the probability is 50% or more, decide it's a cat.
Logistic regression predicts probabilities, and is therefore a regression algorithm. However, it is commonly described as a classification method in the machine learning literature, because it can be (and is often) used to make classifiers. There are also "true" classification algorithms, such as SVM, which only predict an outcome and do not provide a probability. We won't discuss this kind of algorithm here.
Linear vs. Logistic Regression on Classification Problems
As Andrew Ng explains it, with linear regression you fit a polynomial through the data - say, like on the example below we're fitting a straight line through *{tumor size, tumor type}* sample set:



Above, malignant tumors get 1and non-malignant ones get 0, and the green line is our hypothesis $h(x)$. To make predictions we may say that for any given tumor size x, if $h(x)$ gets bigger than 0.5 we predict malignant tumor, otherwise we predict benign.



Q15.
Which of the following statements are true?

Select one or more:
a. None of these.
b. In multiple classes classification One-versus-the-rest method results in ambiguous regions.
c. Goal in classification is to take an input vector x and to assign it to one of K discrete classes.
d. Discriminant function maps each input x, directly onto the class label.
Answer: B, C, D

Linear Models for Classification

- So far we study regression models.
- This lecture discuss an analogous class of models for solving classification problems.
- Goal in classification is to take an input vector **x** and to assign it to one of K discrete classes C_k
  - ✓ Where $k = 1,2,\ldots, K$
- In the most common scenario
  - ✓ the classes are taken to be disjoint
  - ✓ so that each input is assigned to one and only one class
- The input space is thereby divided into *decision regions* whose boundaries are called *decision boundaries* or *decision surfaces*
- We consider linear models for classification
  - ✓ by which we mean that the decision surfaces are linear functions of the input vector **x**
  - ✓ Hence are defined by $(D-1)$-dimensional hyperplanes within the D-dimensional input space

...

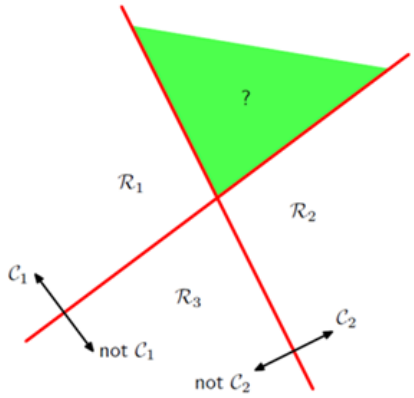
Discriminant Functions

- A **discriminant** is a function that takes an input vector **x** and assigns it to one of K classes, denoted C_k.
- We restrict our attention to *linear discriminants* where the decision surfaces are hyperplanes.
- The simplest representation of a linear discriminant function is obtained by taking a linear function of the input vector so that

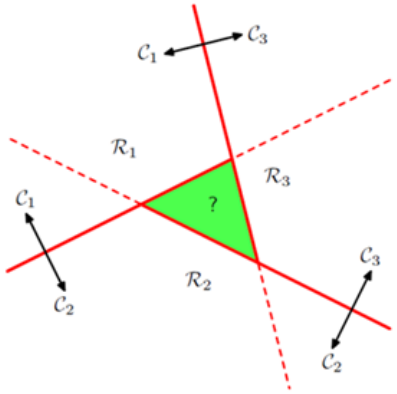
$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

...

Multiple classes classification



One-versus-the-rest method (green region showing the ambiguity)

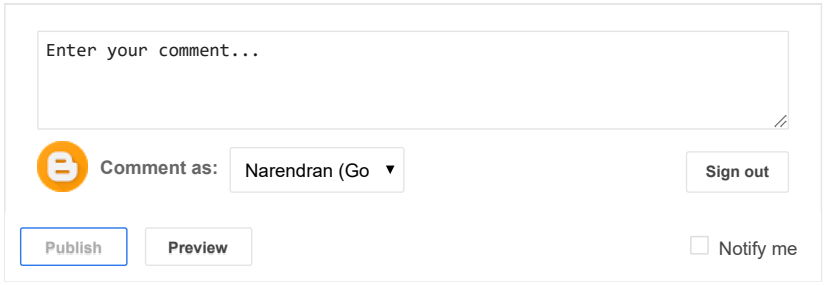


One-versus-one method

...

No comments:

Post a Comment



Home

Subscribe to: Posts (Atom)

Followers

Simple theme. Powered by Blogger.