COMP 562: Introduction to Machine Learning

Lecture 15 : Mid-Term Exam Review

Mahmoud Mostapha

Department of Computer Science

University of North Carolina at Chapel Hill

mahmoudm@cs.unc.edu

October 15, 2018

COMP562 - Lecture 15

Plan for today:

- Topics Covered
- Exam Style
- Sample Questions
- Q&A

Topics Covered

The exam will cover lectures (1-12), topics covered includes:

- Linear Regression
- Logistic Regression
- Regularization
- Generative vs. Discriminative Models
- Naive Bayes Classifier
- Bayesian Networks
- K-means clustering
- Mixture Models

Exam Style

Four questions (20 points) + one extra credit (2 points)

- Question (1) : True or False.[4 points]
- Question (2) : Multiple Choice Questions. [4 points]
- Questions (3,4) : Mathematical Questions (e.g., write formula, derive updates, interpret results,..). [12 points]
- Question (5) : Small extra credit question. [2 points]

Logistics

- In Class Exam: Wednesday, Oct. 17 at 11:15am - 12:30pm
- No electronic devices (e.g., laptops)
- You may need a calculator for some questions
- Open book and notes
  - Think of that as unlimited cheat sheet
  - No need to print every detail in class notes
  - Best practice is to summarize what you learned in class

Advice (for during the exam)

- Solve the easy problems first (e.g. T/F before derivations)
  - If problem seems complicated you're likely missing something
- Dont leave any answer blank!
- if you make an assumption, write it down
- If you look at a question and dont know the answer
  - we probably haven't told you the answer, but we've told you enough to work it out

Sample Questions : True or False

Suppose the dataset in the previous question had 900 cancer-free images and 100 images from cancer patients. If I train a classifier which achieves 85% accuracy on this dataset, it is it a good classifier.

A. True B. False

A football coach whispers a play number n to two players A and B independently. Due to crowd noise, each player imperfectly and independently draws a conclusion about what the play number was. A thinks he heard the number n_A, and B thinks he heard n_B. True or false: n_A and n_B are marginally dependent but conditionally independent given the true play number n.

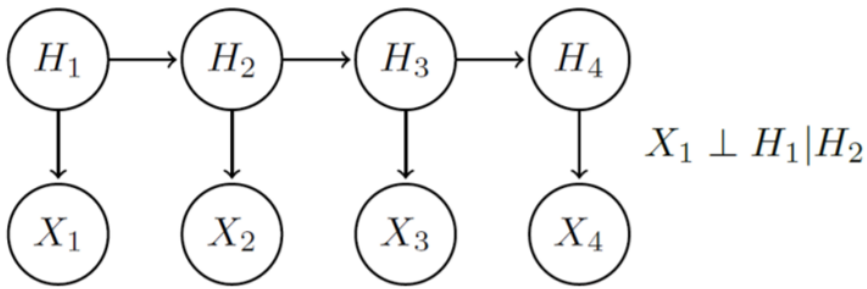
A. True B. False

Sample Questions : True or False

True or False? The EM algorithm finds the global optimum for clustering data from a mixture of Gaussians, assuming the number of clusters K is set to the correct value.

True or False? L_2 penalized linear regression is, in general, more sensitive to outliers than L_1 penalized linear regression. (I.e. one point that is far from the predicted regression line will have more effect on the regression coefficients.)

Does this conditional independence relation hold?



A. True

B. False

Sample Questions : Multiple Choice

Which of the following sentence is FALSE regarding regression?

(A) It relates inputs to outputs.

(B) It is used for prediction.

(C) It may be used for interpretation.

(D) It discovers causal relationships.

For the one-parameter model, mean-Square error (MSE) is defined as follows: $\frac{1}{2N}\sum_{n=1}^{N}(y_n - \beta_0)^2$. We have a half term in the front because,

(A) scaling MSE by half makes gradient descent converge faster.

(B) presence of half makes it easy to do grid search.

(C) it does not matter whether half is there or not.

(D) none of the above

Sample Questions : Multiple Choice

If we we want to encode a true distribution $P(A) = 1/2$, $P(B) = 1/2$, $P(C) = P(D) = 0$ with an estimated distribution $P(A) = 1/2$, $P(B) = 1/4$, $P(C) = P(D) = 1/8$ what is the KL divergence (using log based 2) between them?

(a) 0.5

(b) 1

(c) 4/3

(d) 1.5

(e) none of the above

K-fold cross-validation is

(A) linear in K

(B) quadratic in K

(C) cubic in K

(D) exponential in K

You observe the following while fitting a linear regression to the data: As you increase the amount of training data, the test error decreases and the training error increases. The train error is quite low (almost what you expect it to), while the test error is much higher than the train error.

What do you think is the main reason behind this behavior. Choose the most probable option.

(A) High variance

(B) High model bias

(C) High estimation bias

(D) None of the above

Sample Questions : Derivations

We are going to construct a variant of multi-class logistic regression. This model will have the same feature weights for all classes. However, bias value will vary between classes. Bias for class c will be denoted as α_c

$$p(y = c|\mathbf{x}, \alpha, \beta) = \frac{\exp\left\{-\frac{1}{2}(\beta^T\mathbf{x} - \alpha_c)^2\right\}}{\sum_k \exp\left\{-\frac{1}{2}(\beta^T\mathbf{x} - \alpha_k)^2\right\}}$$

or equivalently

$$p(y|\mathbf{x}, \alpha, \beta) = \prod_{c=1}^{C}\left(\frac{\exp\left\{-\frac{1}{2}(\beta^T\mathbf{x} - \alpha_c)^2\right\}}{\sum_k \exp\left\{-\frac{1}{2}(\beta^T\mathbf{x} - \alpha_k)^2\right\}}\right)^{[y=c]}$$

Sample Questions : Derivations

Write out log-likelihood function for α, β, $\mathcal{LL}(\alpha, \beta; \mathbf{x}, \mathbf{y})$. Hint: use $\mathbf{x}_t$ to denote feature vector of sample t, and y_t to denote label of that sample.

Solution:

$$\mathcal{LL}(\alpha, \beta; \mathbf{x}, \mathbf{y}) = \sum_{t=1}^{T} \sum_{c=1}^{C} [y_t = c] \left(-\frac{1}{2}(\beta^T \mathbf{x}_t - \alpha_c)^2 - \log \sum_k \exp \left\{ -\frac{1}{2}(\beta^T \mathbf{x}_t - \alpha_k)^2 \right\} \right)$$

Write out partial derivatives with respect to α_c.

Solution:

$$\begin{aligned} \frac{\partial}{\partial \alpha_c} \mathcal{LL}(\alpha, \beta; \mathbf{x}, \mathbf{y}) &= \sum_{t=1}^{T} \left([y_t = c](\beta^T \mathbf{x}_t - \alpha_c) - \frac{\exp\left\{-\frac{1}{2}(\beta^T \mathbf{x}_t - \alpha_c)^2\right\}}{\sum_k \exp\left\{-\frac{1}{2}(\beta^T \mathbf{x}_t - \alpha_k)^2\right\}} (\beta^T \mathbf{x}_t - \alpha_c) \right) \\ &= \sum_{t=1}^{T} (\beta^T \mathbf{x}_t - \alpha_c) \left([y_t = c] - \frac{\exp\left\{-\frac{1}{2}(\beta^T \mathbf{x}_t - \alpha_c)^2\right\}}{\sum_k \exp\left\{-\frac{1}{2}(\beta^T \mathbf{x}_t - \alpha_k)^2\right\}} \right) \end{aligned}$$

Sample Questions : Derivations

Find an equation that defines the set of feature vectors $\mathbf{x}$ for which $p(y = 2|\mathbf{x}, \alpha, \beta) = p(y = 3|\mathbf{x}, \alpha, \beta)$. Hint: you do not need information from the previous parts of this problem to answer this question.

Solution:

$$
\begin{aligned}
\frac{\exp\left\{-\frac{1}{2}(\beta^T\mathbf{x} - \alpha_2)^2\right\}}{\sum_k \exp\left\{-\frac{1}{2}(\beta^T\mathbf{x} - \alpha_k)^2\right\}} &= \frac{\exp\left\{-\frac{1}{2}(\beta^T\mathbf{x} - \alpha_3)^2\right\}}{\sum_k \exp\left\{-\frac{1}{2}(\beta^T\mathbf{x} - \alpha_k)^2\right\}} \\
\exp\left\{-\frac{1}{2}(\beta^T\mathbf{x} - \alpha_2)^2\right\} &= \exp\left\{-\frac{1}{2}(\beta^T\mathbf{x} - \alpha_3)^2\right\} \\
(\beta^T\mathbf{x} - \alpha_2)^2 &= (\beta^T\mathbf{x} - \alpha_3)^2 \\
(\beta^T\mathbf{x})^2 - 2\alpha_2(\beta^T\mathbf{x}) + \alpha_2^2 &= (\beta^T\mathbf{x})^2 - 2\alpha_3(\beta^T\mathbf{x}) + \alpha_3^2 \\
2(\alpha_3 - \alpha_2)(\beta^T\mathbf{x}) &= \alpha_3^2 - \alpha 2^2 \\
\beta^T\mathbf{x} &= \frac{\alpha_3 + \alpha_2}{2}
\end{aligned}
$$

Sample Questions : Derivations

Assume we have a random sample that is Bernoulli distributed $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$. We are going to derive the MLE for θ. Recall that a Bernoulli random variable X takes values in $\{0, 1\}$ and has probability mass function given by

$$P(X; \theta) = \theta^X (1 - \theta)^{1-X}.$$

Derive the likelihood, $L(\theta; X_1, \ldots, X_n)$.

Solution:

$$\begin{aligned} L(\theta; X_1, \ldots, X_n) &= \prod_{i=1}^{n} p(X_i; \theta) \\ &= \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}. \end{aligned}$$

Either of the final two steps are acceptable.

Sample Questions : Derivations

Derive the following formula for the log likelihood:

$$\ell(\theta; X_1, \ldots, X_n) = \left(\sum_{i=1}^{n} X_i\right) \log(\theta) + \left(n - \sum_{i=1}^{n} X_i\right) \log(1-\theta).$$

Solution:

$$\begin{aligned} l(\theta; X_1, \ldots, X_n) &= \log L(\theta; X_1, \ldots, X_n) \\ &= \log \left[\theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n - \sum_{i=1}^{n} x_i}\right] \\ &= \left(\sum_{i=1}^{n} x_i\right) \log(\theta) + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-\theta) \end{aligned}$$

Sample Questions : Derivations

Derive the following formula for the MLE: $\hat{\theta} = \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)$.

Solution: To find the MLE we solve $\frac{d}{d\theta}\ell(\theta; X_1, \ldots, X_n) = 0$. The derivative is given by

$$\begin{aligned}\frac{d}{d\theta}\ell(\theta; X_1, \ldots, X_n) &= \frac{d}{d\theta}\left[\left(\sum_{i=1}^{n} x_i\right)\log(\theta) + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-\theta)\right] \\ &= \frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{n - \sum_{i=1}^{n} x_i}{1-\theta}\end{aligned}$$

Next, we solve $\frac{d}{d\theta}\ell(\theta; X_1, \ldots, X_n) = 0$:

$$\begin{aligned}&\frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{n - \sum_{i=1}^{n} x_i}{1-\theta} = 0 \\ \iff &\left(\sum_{i=1}^{n} x_i\right)(1-\theta) - \left(n - \sum_{i=1}^{n} x_i\right)\theta = 0 \\ \iff &\sum_{i=1}^{n} x_i - n\theta = 0 \\ \iff &\hat{\theta} = \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right).\end{aligned}$$

More Questions

Suppose you have regression data generated by a polynomial of degree 3. Characterize the bias-variance of the estimates of the following models on the data with respect to the true model by circling the appropriate entry.

	Bias	Variance
Linear regression	low / high	low / high
Polynomial regression with degree 3	low / high	low / high
Polynomial regression with degree 10	low / high	low / high

Solution:

	Bias	Variance
Linear regression	low / (high)	(low) / high
Polynomial regression with degree 3	(low) / high	(low) / high
Polynomial regression with degree 10	(low) / high	low / (high)

More Questions

You are given a data set of 10,000 students with their sex, height, and hair color. You are trying to build a classifier to predict the sex of a student, so you randomly split the data into a training set and a testing set. Here are the specifications of the data set:

- sex $\in$ {male,female}
- height $\in$ [0,300] centimeters
- hair $\in$ {brown, black, blond, red, green}
- 3240 men in the data set
- 6760 women in the data set

Under the assumptions necessary for Naive Bayes (not the distributional assumptions you might naturally or intuitively make about the dataset) answer each question with **T** or **F** and **provide a one sentence explanation of your answer**:

More Questions

T or F: As height is a continuous valued variable, Naive Bayes is not appropriate since it cannot handle continuous valued variables.

Solution: False. Naive Bayes can handle both continuous and discrete values as long as the appropriate distributions are used for conditional probabilities. For example, Gaussian for continuous and Bernoulli for discrete

T or F: Since there is not a similar number of men and women in the dataset, Naive Bayes will have high test error.

Solution: False. Since the data was randomly split, the same proportion of male and female will be in the training and testing sets. Thus this discrepancy will not affect testing error.

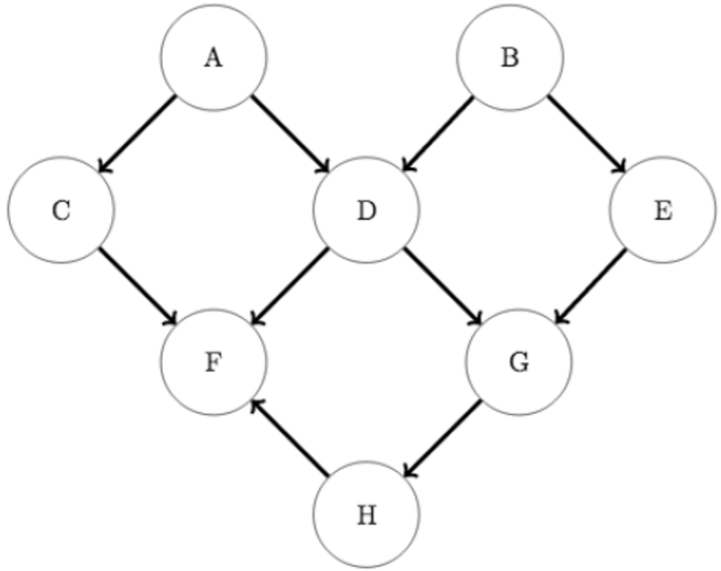
T or F: $P(\texttt{height}|\texttt{sex},\texttt{hair}) = P(\texttt{height}|\texttt{sex})$.

Solution: True. This results from the conditional independence assumption required for Naive Bayes.

T or F: $P(\texttt{height},\texttt{hair}|\texttt{sex}) = P(\texttt{height}|\texttt{sex})P(\texttt{hair}|\texttt{sex})$.

Solution: True. This results from the conditional independence assumption required for Naive Bayes.

More Questions



Consider the Bayesian network in Figure 2. We use $(X \perp\!\!\!\perp Y|Z)$ to denote the fact that X and Y are independent given Z. Answer the following questions:

1. Are there any pairs of point that are independent? If your answer is yes, please list out all such pairs.
   Answer: Yes. (C, E), (C, B), (A, E), (A, B).

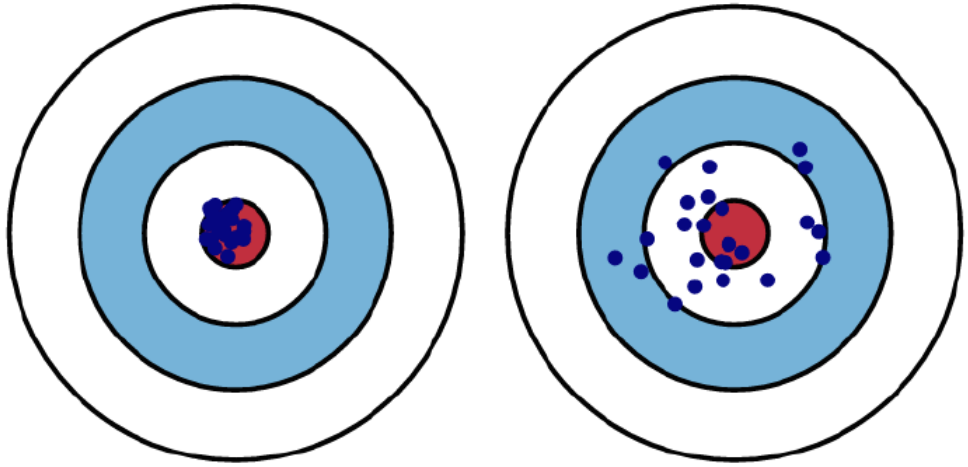
2. Does $(B \perp\!\!\!\perp C|A, D)$ hold? Briefly explain.

   Answer: It holds. We can see that by using the Bayes ball algorithm.

3. Does $(B \perp\!\!\!\perp F|A, D)$ hold? Briefly explain.

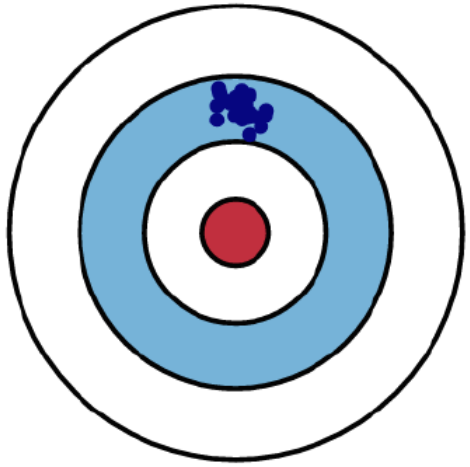
   Answer: It does not hold. By using the Bayes ball algorithm, we can find a path $B \to E \to G \to H \to F$.

More Questions

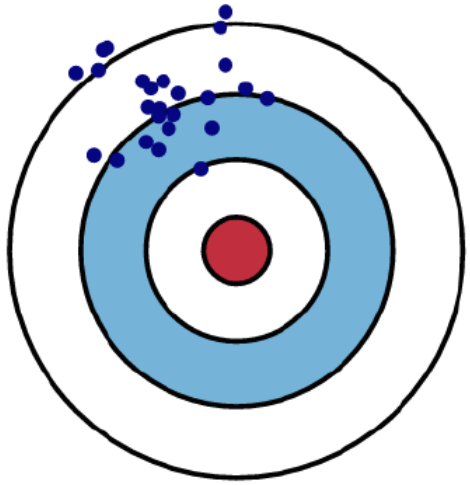


Bias : Low / High
Variance: Low / High

Bias : Low / High
Variance: Low / High



Bias : Low / High
Variance: Low / High



Bias : Low / High
Variance: Low / High

More Questions

Consider the problem of binary classification using the Naive Bayes classifier. You are given two dimensional features (X_1, X_2) and the categorical class conditional distributions in the tables below. The entries in the tables correspond to $P(X_1 = x_1|C_i)$ and $P(X_2 = x_2|C_i)$ respectively. The two classes are *equally likely*.

Class / $X_1 =$	C_1	C_2
-1	0.2	0.3
0	0.4	0.6
1	0.4	0.1

Class / $X_2 =$	C_1	C_2
-1	0.4	0.1
0	0.5	0.3
1	0.1	0.6

Given a data point $(-1, 1)$, calculate the following posterior probabilities:

$P(C_1|X_1 = -1, X_2 = 1) =$ Using Bayes' Rule and conditional independence assumption of Naive Bayes

$$\frac{P(X_1=-1,X_2=1|C_1)P(C_1)}{P(X_1=-1,X_2=1)} = \frac{P(X_1=-1|C_1)P(X_2=1|C_1)P(C_1)}{P(X_1=-1|C_1)P(X_2=1|C_1)P(C_1)+P(X_1=-1|C_2)P(X_2=1|C_2)P(C_2)} = 0.1$$

$P(C_2|X_1 = -1, X_2 = 1) = 1 - P(C_1|X_2 = -1, X_1 = 1) = 0.9$

Q&A

Today

- Mid-Term Exam Review