University of Virginia
Department of Computer Science

CS 6501: Text Mining
Spring 2015

9:30am-9:45am, Thursday, April 9th

Name:
ComputingID:

- This is a **closed book** and **closed notes** quiz. No electronic aids or cheat sheets are allowed.
- There are 2 pages, 3 parts of questions, and 20 total points in this quiz.
- The questions are printed on the **back** of this paper!
- Please carefully read the instructions and questions before you answer them.
- Please pay special attention on your handwriting; if the answers are not recognizable by the instructor, the grading might be inaccurate (*NO* argument about this after the grading is done).
- Try to keep your answers as concise as possible; grading is *not* by keyword matching.

Total	/20

1 True/False Questions (3pts×2)

For the statement you believe it is *False*, please give your brief explanation of it (you do not need to explain anything when you believe it is *True*). *Note the credit can only be granted if your explanation is correct.*

1. $y^* = \arg\max_y p(y|X)$ guarantees the optimal classification boundary in general.
   ***False**, and Explain*: one needs to consider the loss of misclassification for each class.

2. In a 5-fold cross validation, the average F1 measure of classifier 1 is 0.4 and classifier 2 is 0.3. We can confidently conclude that classifier 1 is much better than classifier 2 on this data set.
   ***False**, and Explain*: statistic test is needed to verify if the improvement is significant/meaningful/confident.

2 Multi-choice Questions (4pts×2)

1. Which of the following feature selection algorithms explore class information: (a) (b) (d)
   (a) Information Gain;
   (b) Chi Square;
   (c) Document Frequency;
   (d) Mutual Information.

2. The difference(s) between generative models and discriminative models is(are): (c) (d)
   (a) Discriminative models capture the joint distribution between features and class labels;
   (b) Generative models assume conditional independence among features;
   (c) Generative models can effectively explore unlabeled data;
   (d) Discriminative models provide more flexibility in introducing features.

3 Short Questions (6 pts)

1. Given the following confusion matrix, compute precision, recall and F1 for each class, and the overall accuracy. Element (i, j) in this matrix indicates the instance of class i has been classified to class j.

	1	2	3	R
1	5	1	0	$\frac{5}{6}$
2	2	6	1	$\frac{6}{9}$
3	1	2	8	$\frac{8}{11}$
P	$\frac{5}{8}$	$\frac{6}{9}$	$\frac{8}{9}$	
F	$\frac{5}{7}$	$\frac{6}{9}$	$\frac{8}{10}$	

   Accuracy $= \frac{5+6+8}{5+1+2+6+1+1+2+8}$