Department of Computer Science and Engineering
University of California, San Diego

CSE 291
Spring 2011

Final Examination

Tuesday June 7, 2011

Your name:

Instructions: Answer each question in the space below the question, using the backs of the pages for extra space as necessary. If necessary, you may make assumptions that are reasonable, and that do not make a question trivial. If you do make an assumption, state it clearly. This exam is open-notes. You may use a calculator.

Please read the certification below carefully, and then sign your name to confirm that it is true.

> I certify that I wrote all answers for this exam myself, with no assistance from any other person, and that I spent at most 3 hours total working on this exam.

Signed:

(Question 1) **(Question 1)** [35 points] Suppose that you are working for a phone company that wants to predict churn. Churn is when customers cancel their subscriptions. Let x designate a customer, let *actual* be a binary random variable meaning that the customer actually churns, and let *flag* be a binary random variable meaning that a classifier predicts that the customer will churn. Note that $P(\textit{actual} = 1)$ is the churn base rate, that is the fraction of customers that really do churn. Similarly, $P(\textit{flag} = 1)$ is the fraction of customers that are predicted to churn. We know that the churn base rate is 5%.

Define $a = P(\textit{actual} = 1|\textit{flag} = 1)$ and $b = P(\textit{actual} = 0|\textit{flag} = 0)$. You have hired Dr. Brin as a data mining consultant. He claims that he can train a classifier that will achieve $a = 0.8$ and $b = 0.9$.

(a) [5 points] Write down a confusion matrix and show how to define $P(\textit{actual} = 1)$ and $P(\textit{flag} = 1)$ in terms of the entries in the confusion matrix.

(b) [5 points] Show how to define a and b in terms of the entries in the confusion matrix.

(c) [5 points] Explain in English the meaning of Dr. Brin's claim. Would such a classifier be useful for making decisions, or not?

(d) [10 points] Prove mathematically that Dr. Brin is wrong, because no such classifier exists.

(e) [10 points] Provide a precise but intuitive and simple argument that confirms that no such classifier exists.

(Question 2) [39 points] For each statement below, clearly write "True" if it is mostly true, or "False" if it is mostly false. Then in the space below, write one or two sentences explaining why or how the statement is true or false. The maximum score for each answer is three points.

All statements below are based on the following scenario. You are training a two-class linear SVM classifier. You have selected a number of informative features. Accuracy on the training set is b, while accuracy on an independent test set is a. The difference is $d = b - a$.

Assume that a is too low to be useful for the intended application. Your goal is to make the classifier more useful by making a higher.

1. Typically (i.e. under normal circumstances), $d > 0$.

   From now on, assume that circumstances are normal and that d has the appropriate sign. For the following statements, assume that $|d|$ is large.

2. Large $|d|$ is a strong indication of overfitting.

3. Using twice as many features is likely to improve a.

4. Using twice as many training examples is likely to improve a.

5. Using a nonlinear kernel is likely to improve a.

6. Using stronger regularization is likely to improve a.

7. Stronger regularization means a larger value for the SVM algorithm parameter C.

For the following statements, assume that $|d|$ is small, i.e. training error b and test error a are almost equal.

8. Using twice as many features is likely to improve a.

9. Using twice as many training examples is likely to improve a.

10. Using a nonlinear kernel is likely to improve a.

11. Using stronger regularization is likely to improve a.

12. Increasing the number of digits of precision of feature values is likely to improve a.

13. Doubling the number of features by including the square of each existing feature is likely to improve a.

(Question 3) [50 points] Suppose that you want to use reinforcement learning to optimize the design of a website such as Facebook. Your goal is to maximize the total length of time that a user spends on the site.

(a) [5 points] One user is one random instance of the environment. What is a reasonable state vector to represent a user?

(b) [5 points] The agent is the website. What are some reasonable actions available to the agent?

(c) [5 points] A Markov decision process (MDP) is a four-tuple $\langle S, A, p, r \rangle$. Explain what S, A, and p are in this application.

(d) [5 points] Give a reasonable definition of r in this application that is explicit and precise.

(e) [5 points] For offline learning, what is the format of training data needed?

(f) [5 points] Explain how the company can acquire appropriate training data.

(g) [5 points] Explain how the company can use linear regression to learn a Q function.

(h) [5 points] Explain how the company can use the Q function to make decisions with respect to a future user.

(i) [5 points] Describe a sensible way for the company to measure empirically the effectiveness of reinforcement learning.

(j) [5 points] The approach suggested above has never been implemented yet. What do you think are the most significant reasons why it might fail?

(Question 4) [20 points] In a famous essay entitled *On Bullshit*, the philosopher Harry G. Frankfurt wrote

> It is just this lack of connection to a concern with truth—this indifference to how things really are–that I regard as of the essence of bullshit.

Does data mining lack a concern with truth? In your answer, discuss both algorithms and applications.

Another quotation from the essay that is relevant is

> The realms of advertising and of public relations, and the nowadays closely related realm of politics, are replete with instances of bullshit so unmitigated that they can serve among the most indisputable and classic paradigms of the concept. And in these realms there are exquisitely sophisticated craftsmen who—with the help of advanced and demanding techniques of market research, of public opinion polling, of psychological testing, and so forth—dedicate themselves tirelessly to getting every word and image they produce exactly right.