CSE 446 Midterm Exam, Spring 2013

1. Personal info:
   - Name:
   - Student ID:
   - E-mail address:
2. There should be 9 numbered pages in this exam (including this cover sheet).
3. You can use any material you brought: any book, class notes, your print outs of class materials that are on the class website, including my annotated slides and relevant readings. You cannot use materials brought by other students. Calculators are not necessary. Laptops, PDAs, phones and Internet access are not allowed.
4. If you need more room to work out your answer to a question, use the back of the page and clearly mark on the front of the page if we are to look at what's on the back.
5. Work efficiently. Some questions are easier, some more difficult. Be sure to give yourself time to answer all of the easy ones, and avoid getting bogged down in the more difficult ones before you have answered the easier ones.
6. You have 50 minutes.
7. Good luck!

Question	Topic	Max. score	Score
1	True/False	20	
2	Short Answer	24	
3	Maximum Likelihood	13	
4	Decision Trees	18	
5	Decision Boundaries	25	

1 [20 points, 2 points each] True/False (Please add a 1 sentence justification.)

1. **true/false** the following model can be learned by linear regression: $y_i = e^{\beta x_i + \epsilon_i}$ where $\epsilon_i \sim N(0, \sigma^2)$ is iid Gaussian noise.

2. **true/false** When we have less data and the model complexity stays the same, overfitting is more likely.

3. **true/false** When our data points have fewer features, overfitting is more likely.

4. **true/false:** As linear regression is given more and more data, its training error will eventually approach 0 (assuming there is no noise in the data).

5. **true/false:** In machine learning, the bias is always a bigger source of error than the variance.

6. **true/false:** As the number of iterations goes to infinity, boosting is always guaranteed reach zero training error.

7. **true/false:** Suppose a dataset is linearly separable, a logistic regressor with regularization parameter $\lambda > 0$ is guaranteed to separate the data.

8. **true/false:** Nearest neighbors is more efficient at training time than logistic regression.

9. **true/false:** Nearest neighbors is more efficient at test time than logistic regression.

10. (circle the correct answers) Underfitting is generally a symptom of high **bias / variance**, while overfitting is generally a symptom of high **bias / variance**

2 [24 points] Short Answer

1. [4 points] In ordinary least squares regression under what assumption is the coefficient estimator $\hat{w} = (X^T X)^{-1} X^T Y$ that minimizes the residual sum of squares also the maximum likelihood estimator? (No proof required).

2. [4 points] Briefly explain the difference between (batch) gradient descent and stochastic gradient descent. Give an example of when you might prefer one over the other.

3. [4 points] Give 3 examples of different types of classifiers we have covered so far.

4. [4 points] Why do you use kernels to model a projection from attributes into a feature space, instead of simply projecting the dataset directly and using this as input into our learning algorithm?

5. Suppose you wanted to use a regularization function that penalizes parameters exponentially. In particular, we minimize:

$$\sum_{j=1}^{N} \ell(\mathbf{w}, \mathbf{x}^j) + \lambda \sum_i e^{w_i}.$$

- [4 points] Provide a simple stochastic gradient descent update for this objective function, if you are using hinge loss for $\ell(\mathbf{w}, \mathbf{x})$. (In class, we discussed two definitions of the hinge loss, you can use either here.)

- [4 points] What's a potential issue with this type of regularization, if the parameters can take negative values?

3 [13 points] Maximum Likelihood

Suppose you have n IID sample data points $\{x_1, ..., x_n\}$. These data points come from a distribution where the probability of a given datapoint x is $P(x) = \frac{1}{\theta}e^{-\frac{1}{\theta}x}$. Prove that the MLE estimate of parameter θ is the sample mean.

4 [18 points, 6 points each] Decision Trees

For this problem you will respond to several questions concerning the dataset shown below. You will be using a decision tree to classify whether or not an advertisement was clicked based on its size, position, and whether or not it played a sound.

Clicked	Size	Position	Sound
F	Big	Top	No
F	Small	Middle	Yes
F	Small	Middle	Yes
T	Small	Bottom	No
T	Big	Bottom	No
F	Big	Top	Yes
T	Big	Bottom	Yes
T	Small	Middle	No
T	Small	Middle	No
F	Big	Top	No

1. What is the initial entropy of *Clicked*?

2. Assume that *Position* is chosen for the root of the decision tree. What is the information gain associated with this attribute?

3. Draw the full decision tree learned from this data (without any pruning).

5 [25 points, 5 points each] Decision Boundaries

For each of the settings below, show the data points as pluses and minus, and draw the best decision boundary for each method.

1. Draw a 2D dataset where 1-NN will perform better than SVMs with a linear kernel, even with the best choice of regularization.

2. Draw a 2D dataset where 1-NN will perform worse than SVMs with a linear kernel, with the best choice of regularization.

3. Draw a 2D dataset where SVMs with linear kernels will perform poorly, but second-order polynomial kernels will give perfect classification.

4. Draw a 2D dataset where decision trees of depth 2 will perform poorly, but depth 3 will be exact. (In continuous spaces, decision trees don't split on all possible values, like with discrete attributes. Instead, we pick an attribute x_i and a threshold t and split on one side for $x_i > t$ and on the other for $x_i \leq t$. Here, we can split on the same variable multiple times in the construction of the tree, but using different threshold values.)

5. Draw an infinite 2D dataset where decision tree with bounded depth will perform poorly, a simple SVM with linear kernels will provide perfect classification. (Use the same types of decision trees as in the previous question.)