10-601 Machine Learning, Fall 2009: Midterm

Monday, November 2nd—2 hours

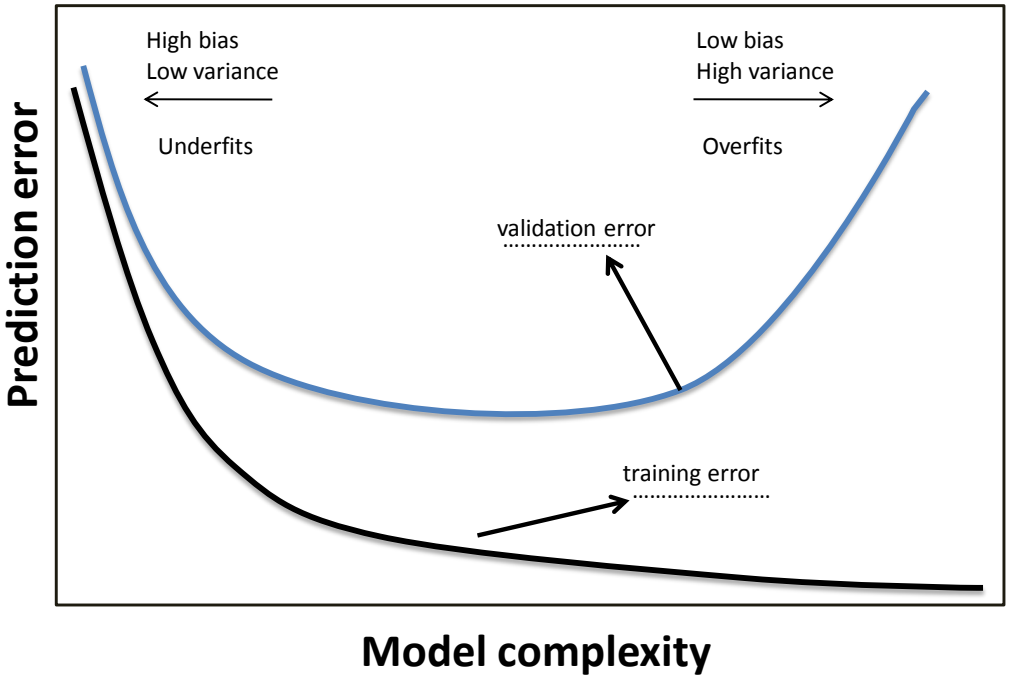
---

1. Personal info:
   - Name:
   - Andrew account:
   - E-mail address:
2. You are permitted two pages of notes and a calculator. Please turn off all cell phones and other noisemakers.
3. There should be 26 numbered pages in this exam (including this cover sheet). If the last page is not numbered 26 please let us know immediately. The exam is "thick" because we provided extra space between each question. If you need additional paper please let us know.
4. There are 13 questions worth a total of 154 points (plus some extra credit). 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.
5. There are extra-credit questions at the end. The grade curve will be made without considering extra credit. Then we will use the extra credit to try to bump your grade up without affecting anyone else's.
6. You have 120 minutes. Good luck!

Question	Topic	Max. score	Score
1	Training and Validation	8	
2	Bias and Variance	6	
3	Experimental Design	16	
4	Logistic Regression	8	
5	Regression with Regularization	10	
6	Controlling Over-Fitting	6	
7	Decision Boundaries	12	
8	k-Nearest Neighbor Classifier	6	
9	Decision Trees	16	
10	Principal Component Analysis	12	
11	Bayesian Networks	30	
12	Graphical Model Inference	8	
13	Gibbs Sampling	16	
	Total	154	
14	Extra Credit	22	

1 Training and Validation [8 Points]

The following figure depicts training and validation curves of a learner with increasing model complexity.



1. [**Points: 2 pts**] Which of the curves is more likely to be the training error and which is more likely to be the validation error? Indicate on the graph by filling the dotted lines.

2. [**Points: 4 pts**] In which regions of the graph are bias and variance low and high? Indicate clearly on the graph with four labels: "low variance", "high variance", "low bias", "high bias".

3. [**Points: 2 pts**] In which regions does the model overfit or underfit? Indicate clearly on the graph by labeling "overfit" and "underfit".

2 Bias and Variance [6 Points]

A set of data points is generated by the following process: $Y = w_0 + w_1X + w_2X^2 + w_3X^3 + w_4X^4 + \epsilon$, where X is a real-valued random variable and ϵ is a Gaussian noise variable. You use two models to fit the data:

Model 1: $Y = aX + b + \epsilon$

Model 2: $Y = w_0 + w_1X^1 + ... + w_9X^9 + \epsilon$

1. [**Points: 2 pts**] Model 1, when compared to Model 2 using a fixed number of training examples, has a *bias* which is:
   (a) Lower
   (b) Higher ★
   (c) The Same
2. [**Points: 2 pts**] Model 1, when compared to Model 2 using a fixed number of training examples, has a *variance* which is:
   (a) Lower ★
   (b) Higher
   (c) The Same
3. [**Points: 2 pts**] Given 10 training examples, which model is more likely to overfit the data?
   (a) Model 1
   (b) Model 2 ★

★ **SOLUTION:** Correct answers are indicated with a star next to them.

3 Experimental design [16 Points]

For each of the listed descriptions below, circle whether the experimental set up is *ok* or *problematic*. If you think it is problematic, briefly state **all** the problems with their approach:

1. **[Points: 4 pts]** A project team reports a low training error and claims their method is good.

   (a) Ok

   (b) Problematic ★

   ★ **SOLUTION:** Problematic because training error is an optimistic estimator of test error. Low training error does not tell much about the generalization performance of the model. To prove that a method is good they should report their error on independent test data.

2. **[Points: 4 pts]** A project team claimed great success after achieving 98 percent classification accuracy on a binary classification task where one class is very rare (e.g., detecting fraud transactions). Their data consisted of 50 positive examples and 5 000 negative examples.

   (a) Ok

   (b) Problematic ★

   ★ **SOLUTION:** Think of classifier which predicts everything as the majority class. The accuracy of that classifier will be 99%. Therefore 98% accuracy is not an impressive result on such an unbalanced problem.

3. **[Points: 4 pts]** A project team split their data into training and test. Using their training data and cross-validation, they chose the best parameter setting. They built a model using these parameters and their training data, and then report their error on test data.

   (a) Ok ★

   (b) Problematic

   ★ **SOLUTION:** OK.

4. **[Points: 4 pts]** A project team performed a feature selection procedure on the full data and reduced their large feature set to a smaller set. Then they split the data into test and training portions. They built their model on training data using several different model settings, and report the the best test error they achieved.

   (a) Ok

   (b) Problematic ★

   ★ **SOLUTION:** Problematic because:

   (a) Using the full data for feature selection will leak information from the test examples into the model. The feature selection should be done exclusively using training and validation data not on test data.

   (b) The best parameter setting should not be chosen based on the test error; this has the danger of overfitting to the test data. They should have used validation data and use the test data only in the final evaluation step.

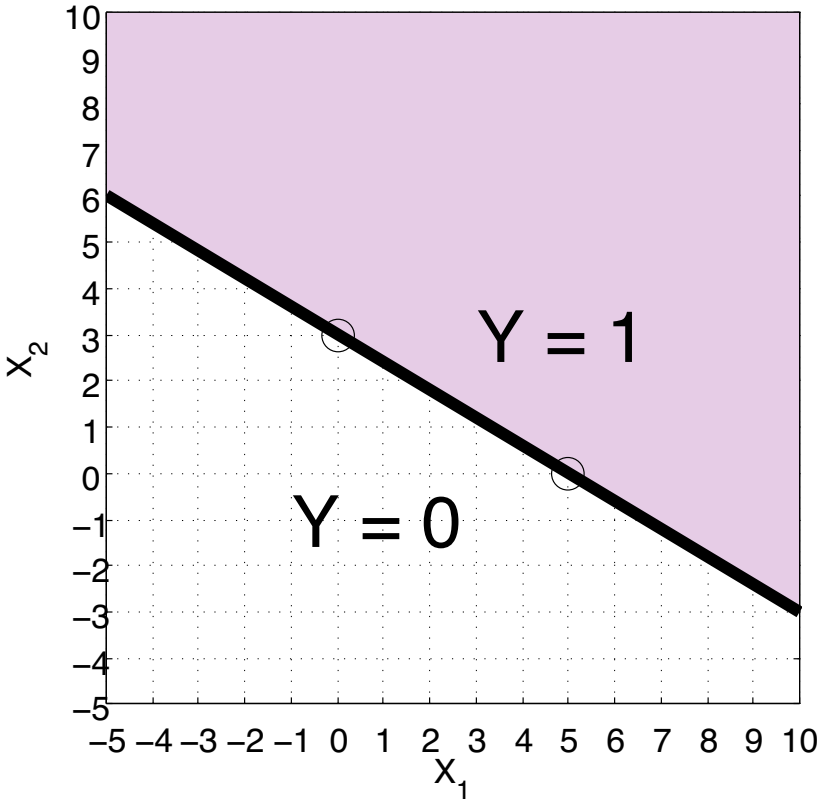
4 Logistic Regression [8 Points]

Suppose you are given the following classification task: predict the target $Y \in \{0, 1\}$ given two real valued features $X_1 \in \mathbb{R}$ and $X_2 \in \mathbb{R}$. After some training, you learn the following decision rule:

Predict $Y = 1$ **iff** $w_1 X_1 + w_2 X_2 + w_0 \geq 0$ **and** $Y = 0$ **otherwise**

where $w_1 = 3$, $w_2 = 5$, and $w_0 = -15$.

1. **[Points: 6 pts]** Plot the decision boundary and label the region where we would predict $Y = 1$ and $Y = 0$.



★ **SOLUTION:** See above figure.

2. **[Points: 2 pts]** Suppose that we learned the above weights using logistic regression. Using this model, what would be our prediction for $P(Y = 1 \mid X_1, X_2)$? (You may want to use the sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$.)

$$\mathbf{P}\left(Y = 1 \mid X_1, X_2\right) =$$

★ **SOLUTION:**

$$\mathbf{P}\left(Y = 1 \mid X_1, X_2\right) = \frac{1}{1 + \exp^{-(3X_1 + 5X_2 - 15)}}$$

5 Regression with Regularization [10 Points]

You are asked to use regularized linear regression to predict the target $Y \in \mathbb{R}$ from the eight-dimensional feature vector $X \in \mathbb{R}^8$. You define the model $Y = w^T X$ and then you recall from class the following three objective functions:

$$\min_w \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 \tag{5.1}$$

$$\min_w \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 \quad + \quad \lambda \sum_{j=1}^{8} w_j^2 \tag{5.2}$$

$$\min_w \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 \quad + \quad \lambda \sum_{j=1}^{8} |w_j| \tag{5.3}$$

1. **[Points: 2 pts]** Circle regularization terms in the objective functions above.

   ★ **SOLUTION:** The regularization term in 5.2 is $\lambda \sum_{j=1}^{8} w_j^2$ and in 5.3 is $\lambda \sum_{j=1}^{8} |w_j|$.

2. **[Points: 2 pts]** For large values of λ in objective 5.2 the bias would:

   (a) increase ★
   (b) decrease
   (c) remain unaffected

3. **[Points: 2 pts]** For large values of λ in objective 5.3 the variance would:

   (a) increase
   (b) decrease ★
   (c) remain unaffected

4. **[Points: 4 pts]** The following table contains the weights learned for all three objective functions (not in any particular order):

	Column A	Column B	Column C
w_1	0.60	0.38	0.50
w_2	0.30	0.23	0.20
w_3	-0.10	-0.02	0.00
w_4	0.20	0.15	0.09
w_5	0.30	0.21	0.00
w_6	0.20	0.03	0.00
w_7	0.02	0.04	0.00
w_8	0.26	0.12	0.05

   Beside each objective write the appropriate column label (A, B, or C):

   - Objective 5.1: ★ **Solution:** A
   - Objective 5.2: ★ **Solution:** B
   - Objective 5.3: ★ **Solution:** C

6 Controlling Overfitting [6 Points]

We studied a number of methods to control overfitting for various classifiers. Below, we list several classifiers and actions that might affect their bias and variance. Indicate (by circling) how the bias and variance change in response to the action:

1. [**Points: 2 pts**] Reduce the number of leaves in a decision tree:

   ★ **SOLUTION:**

Bias	Variance
Decrease	Decrease ★
★ Increase	Increase
No Change	No Change

2. [**Points: 2 pts**] Increase k in a k-nearest neighbor classifier:

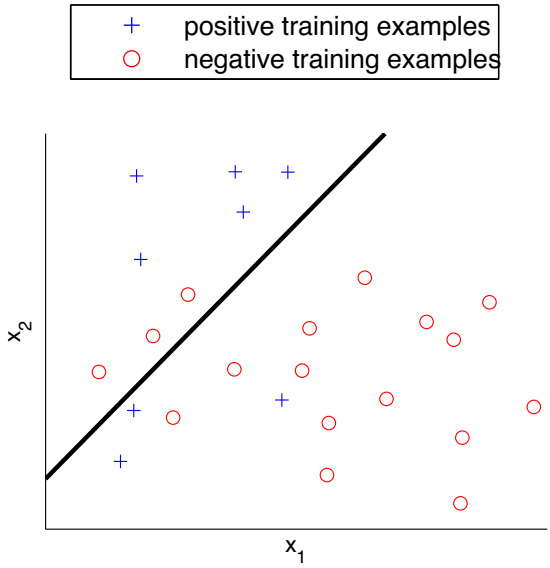
Bias	Variance
Decrease	Decrease ★
★ Increase	Increase
No Change	No Change

3. [**Points: 2 pts**] Increase the number of training examples in logistic regression:

Bias	Variance
Decrease	Decrease ★
Increase	Increase
★ No Change	No Change

7 Decision Boundaries [12 Points]

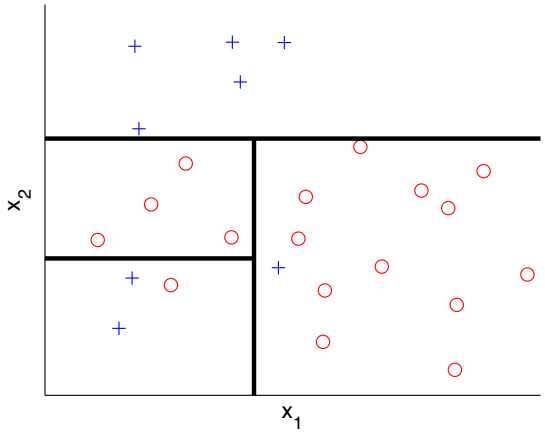
The following figures depict decision boundaries of classifiers obtained from three learning algorithms: decision trees, logistic regression, and nearest neighbor classification (in some order). Beside each of the three plots, write the **name** of the learning algorithm and the **number of mistakes** it makes on the training data.



[Points: 4 pts]

Name: ★ Logistic regression

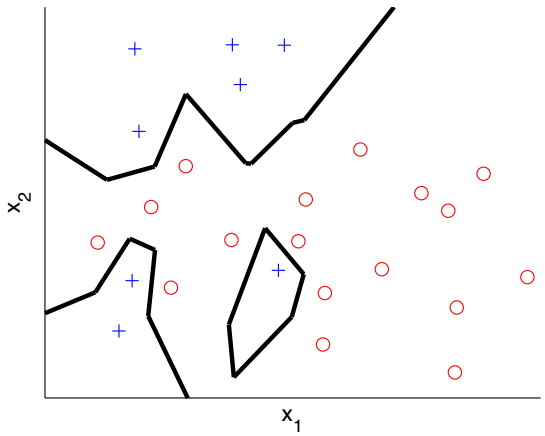
Number of mistakes: ★ 6



[Points: 4 pts]

Name: ★ Decision tree

Number of mistakes: ★ 2



[Points: 4 pts]

Name: ★ k-nearest neighbor

Number of mistakes: ★ 0

11 Bayesian Networks [30 Points]

Consider the Bayes net:

$$H \rightarrow U \leftarrow P \leftarrow W$$

Here, $H \in \{T, F\}$ stands for "10-601 homework due tomorrow"; $P \in \{T, F\}$ stands for "mega-party tonight"; $U \in \{T, F\}$ stands for "up late"; and $W \in \{T, F\}$ stands for "it's a weekend."

1. **[Points: 6 pts]** Which of the following conditional or marginal independence statements follow from the above network structure? Answer *true* or *false* for each one.
   (a) $H \perp P$ ★ **Solution:** *True*
   (b) $W \perp U \mid H$ ★ **Solution:** *False*
   (c) $H \perp P \mid U$ ★ **Solution:** *False*
2. **[Points: 4 pts]** *True* or *false*: Given the above network structure, it is possible that $H \perp U \mid P$. Explain briefly.

   ★ **SOLUTION:** True. This can be achieved through context specific independence (CSI) or accidental independence.
3. **[Points: 4 pts]** Write the joint probability of H, U, P, and W as the product of the conditional probabilities described by the Bayesian Network:

   ★ **SOLUTION:** The joint probability can be written as:

   $$\mathbf{P}(H, U, P, W) = \mathbf{P}(H)\,\mathbf{P}(W)\,\mathbf{P}(P \mid W)\,\mathbf{P}(U \mid H, P)$$
4. **[Points: 4 pts]** How many independent parameters are needed for this Bayesian Network?

   ★ **SOLUTION:** The network will need 8 independent parameters:
   - $\mathbf{P}(H)$: 1
   - $\mathbf{P}(W)$: 1
   - $\mathbf{P}(P \mid W)$: 2
   - $\mathbf{P}(U \mid H, P)$: 4
5. **[Points: 2 pts]** How many independent parameters would we need if we made *no* assumptions about independence or conditional independence?

   ★ **SOLUTION:** A model which makes no conditional independence assumptions would need $2^4 - 1 = 15$ parameters.

6. **[Points: 10 pts]** Suppose we observe the following data, where each row corresponds to a single observation, i.e., a single evening where we observe all 4 variables:

H	U	P	W
F	F	F	F
T	T	F	T
T	T	T	T
F	T	T	T

Use Laplace smoothing to estimate the parameters for each of the conditional probability tables. Please write the tables in the following format:

$\mathbf{P}(Y = T) = 2/3$

Y	Z	$\mathbf{P}(X = T \mid Y, Z)$
T	T	$1/3$
T	F	$3/4$
F	T	$1/8$
F	F	0

(If you prefer to use a calculator, please use decimals with at least three places after the point.)

★ **SOLUTION:** The tables are:

$$\mathbf{P}(H = T) = \frac{2+1}{4+2} = \frac{1}{2} \qquad \mathbf{P}(W = T) = \frac{3+1}{4+2} = \frac{2}{3}$$

W	$\mathbf{P}(P = T \mid W)$
T	$\frac{2+1}{3+2} = \frac{3}{5}$
F	$\frac{0+1}{1+2} = \frac{1}{3}$

H	P	$\mathbf{P}(X = T \mid H, P)$
T	T	$\frac{1+1}{1+2} = \frac{2}{3}$
T	F	$\frac{1+1}{1+2} = \frac{2}{3}$
F	T	$\frac{1+1}{1+2} = \frac{2}{3}$
F	F	$\frac{0+1}{1+2} = \frac{1}{3}$

14.2 Make your Own Question [8 Points]

1. [**Points: 4 pts**] Writing interesting machine learning questions is difficult. Write your own question about material covered 10-601. You will get maximum credit for writing an interesting and insightful question.

2. [**Points: 4 pts**] Attempt to answer your question. You will get maximum credit for providing an insightful (and correct!) answer.