10-601 Machine Learning, Fall 2009: Midterm

Monday, November 2$^{\text{nd}}$—2 hours

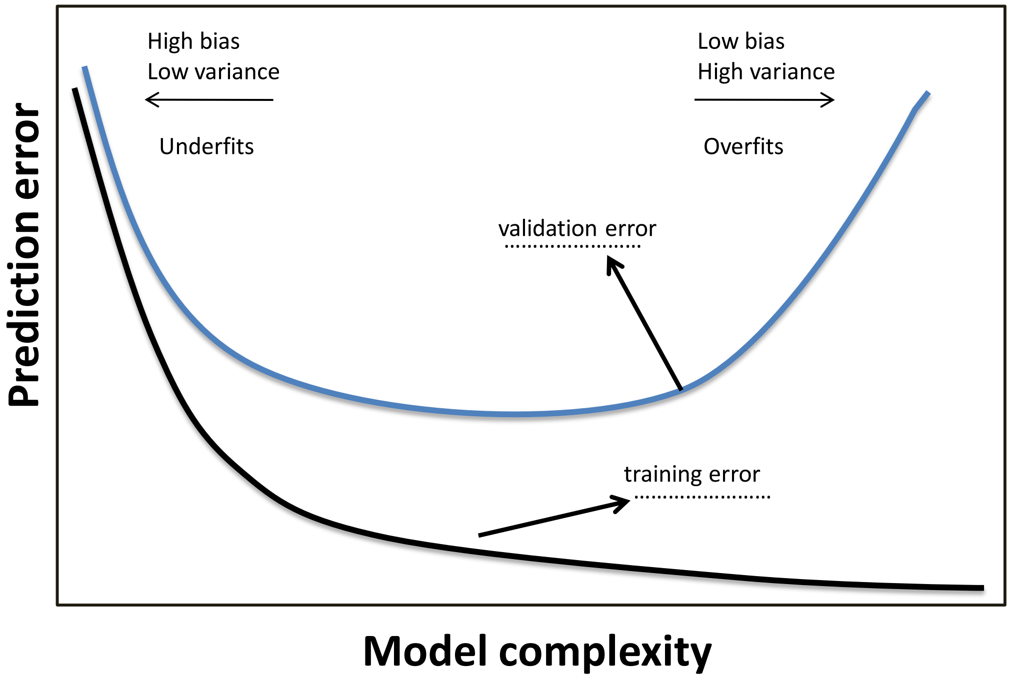
---

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: $Y = w_0 + w_1 X + w_2 X^2 + w_3 X^3 + w_4 X^4 + \epsilon$, 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: $Y = w_0 + w_1 X^1 + ... + w_9 X^9 + \epsilon$

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.

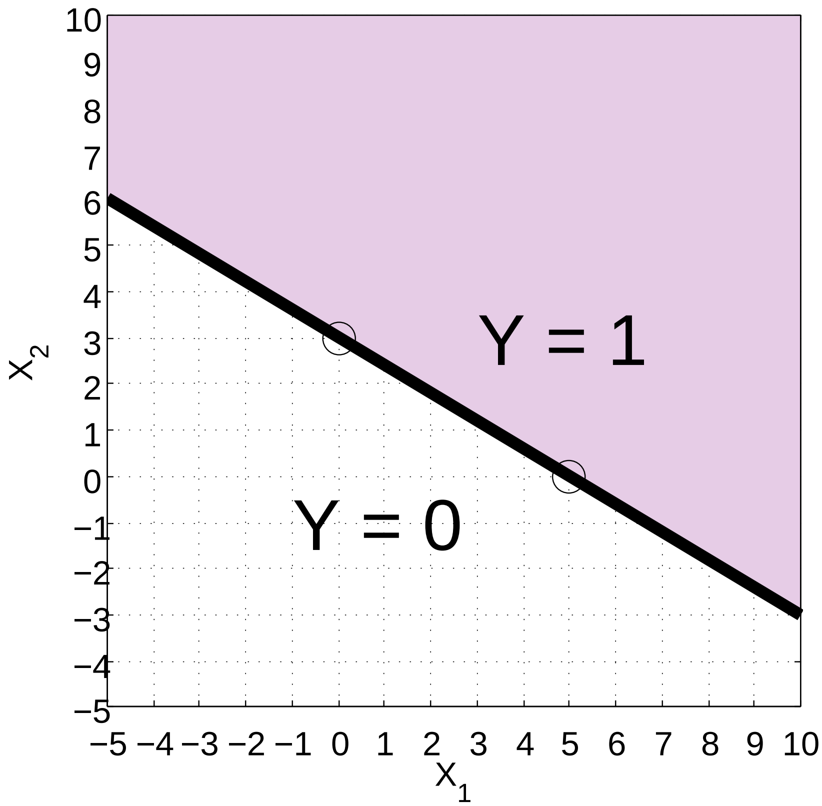
4 Logistic Regression [8 Points]

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:

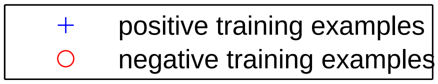
Bias	Variance
Decrease	Decrease ★
★ Increase	Increase
No Change	No Change

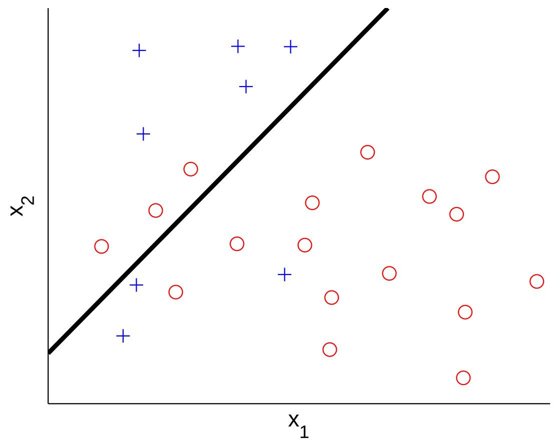
3. **[Points: 2 pts]** Increase the number of training examples in logistic regression:

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.

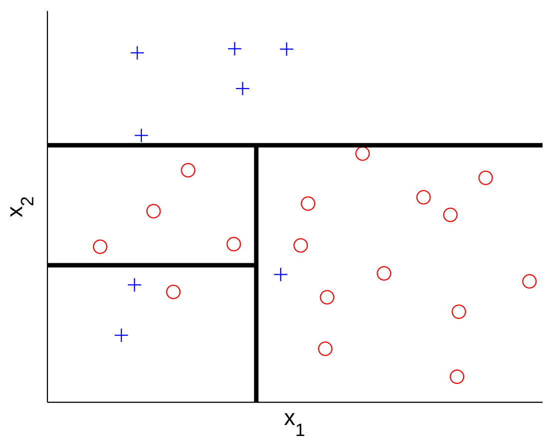




[Points: 4 pts]

Name: ★ Logistic regression

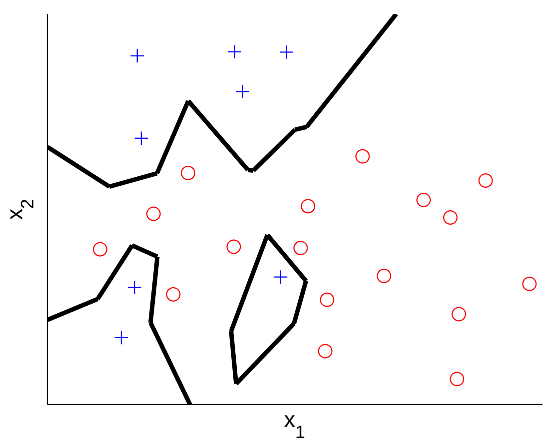
Number of mistakes: ★ 6



[Points: 4 pts]

Name: ★ Decision tree

Number of mistakes: ★ 2



[Points: 4 pts]

Name: ★ k-nearest neighbor

Number of mistakes: ★ 0

8 k-Nearest Neighbor Classifiers [6 Points]

In Fig. 1 we depict training data and a single test point for the task of classification given two continuous attributes X_1 and X_2. For each value of k, circle the label predicted by the k-nearest neighbor classifier for the depicted test point.

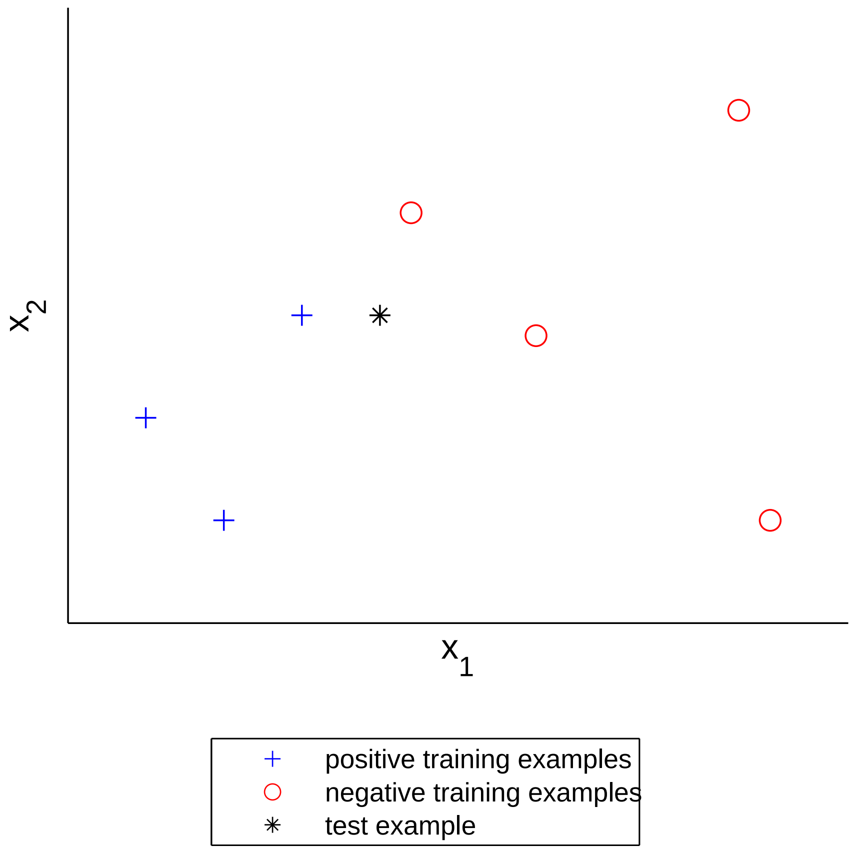


Figure 1: Nearest neighbor classification

1. **[Points: 2 pts]** Predicted label for $k = 1$:

   (a) positive ★ **(b)** negative

2. **[Points: 2 pts]** Predicted label for $k = 3$:

   (a) positive **(b)** negative ★

3. **[Points: 2 pts]** Predicted label for $k = 5$:

   (a) positive ★ **(b)** negative

9 Decision Trees [16 Points]

Suppose you are given six training points (listed in Table 1) for a classification problem with two binary attributes X_1, X_2, and three classes $Y \in \{1, 2, 3\}$. We will use a decision tree learner based on information gain.

X_1	X_2	Y
1	1	1
1	1	1
1	1	2
1	0	3
0	0	2
0	0	3

Table 1: Training data for the decision tree learner.

1. **[Points: 12 pts]** Calculate the information gain for both X_1 and X_2. You can use the approximation $\log_2 3 \approx 19/12$. Report information gains as fractions or as decimals with the precision of three decimal digits. Show your work and circle your final answers for $\text{IG}(X_1)$ and $\text{IG}(X_2)$.

★ **SOLUTION:** The equation for information gain, entropy, and conditional entropy are given by (respectively):

$$\begin{aligned}
\mathsf{IG}(X) &= \mathsf{H}(Y) - \mathsf{H}(Y \mid X) \\
\mathsf{H}(X) &= -\sum_x \mathbf{P}(X = x) \log_2 \mathbf{P}(X = x) \\
\mathsf{H}(Y \mid X) &= \sum_x \mathbf{P}(X = x) \sum_y \mathbf{P}(Y = y \mid X = x) \log_2 \mathbf{P}(Y = y \mid X = x)
\end{aligned}$$

Using these equations we can derive the information gain for each split. First we compute the entropy $\mathsf{H}(Y)$:

$$\begin{aligned}
\mathsf{H}(Y) &= -\sum_{y_i=1}^{n=3} \mathbf{P}(Y = y_i) \log_2 \mathbf{P}(Y = y_i) \\
&= -\sum_{y_i=1}^{n=3} \frac{1}{3} \log_2 \frac{1}{3} = \log_2 3 \approx \frac{19}{12}
\end{aligned}$$

For the X_1 split we compute the conditional entropy:

$$\begin{aligned}
\mathsf{H}(Y \mid X_1) &= -\mathbf{P}(X_1 = 0) \sum_{y_i=1}^{n=3} \mathbf{P}(Y = y_i \mid X_1 = 0) \log_2 \mathbf{P}(Y = y_i \mid X_1 = 0) \quad + \\
&\quad -\mathbf{P}(X_1 = 1) \sum_{y_i=1}^{n=3} \mathbf{P}(Y = y_i \mid X_1 = 1) \log_2 \mathbf{P}(Y = y_i \mid X_1 = 1) \\
&= -\left[\frac{2}{6}\left(\frac{0}{2}\log_2\frac{0}{2} + \frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{4}{6}\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4}\right)\right] \\
&= -\left(-\frac{2}{6} - 1\right) \\
&= \frac{4}{3}
\end{aligned}$$

Similarly for the X_2 split we compute the conditional entropy:

$$\begin{aligned}
\mathsf{H}(Y \mid X_2) &= -\mathbf{P}(X_2=0)\sum_{y_i=1}^{n=3}\mathbf{P}(Y=y_i \mid X_2=0)\log_2\mathbf{P}(Y=y_i \mid X_2=0) \quad + \\
&\quad -\mathbf{P}(X_2=1)\sum_{y_i=1}^{n=3}\mathbf{P}(Y=y_i \mid X_2=1)\log_2\mathbf{P}(Y=y_i \mid X_2=1) \\
&= -\left[\frac{3}{6}\left(\frac{0}{3}\log_2\frac{0}{3}+\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3}\right)+\frac{3}{6}\left(\frac{2}{3}\log_2\frac{2}{3}+\frac{1}{3}\log_2\frac{1}{3}+\frac{0}{3}\log_2\frac{0}{3}\right)\right] \\
&\approx -\left(\frac{2}{3}-\frac{19}{12}\right) \\
&= \frac{11}{12}
\end{aligned}$$

The final information gain for each split is then:

$$\begin{aligned}
\mathsf{IG}(X_1) = \mathsf{H}(Y) - \mathsf{H}(Y \mid X_1) \approx \frac{19}{12}-\frac{4}{3}=\frac{3}{12}=\frac{1}{4} \\
\mathsf{IG}(X_2) = \mathsf{H}(Y) - \mathsf{H}(Y \mid X_2) \approx \frac{19}{12}-\frac{11}{12}=\frac{8}{12}=\frac{2}{3}
\end{aligned}$$

2. **[Points: 4 pts]** Report which attribute is used for the first split. Draw the decision tree resulting from using this split alone. Make sure to label the split attribute, which branch is which, and what the predicted label is in each leaf. How would this tree classify an example with $X_1 = 0$ and $X_2 = 1$?

★ **SOLUTION:** Since the information gain of X_2 is greater than X_1's information gain, we choose to split on X_2. See the resulted decision tree in Fig. 2. An example with $X_1 = 0$ and $X_2 = 1$ will be classified as $Y = 1$ on this tree since $X_2 = 1$.

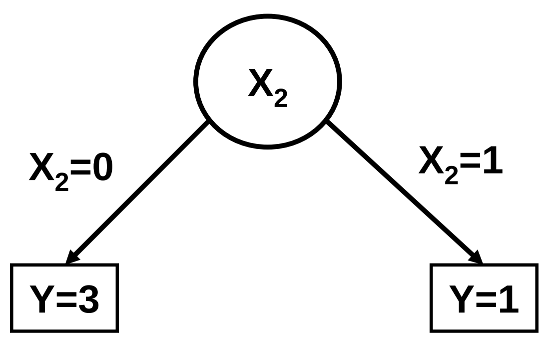


Figure 2: The decision tree for question 9.2

10 Principal Component Analysis [12 Points]

Plotted in Fig. 3 are two dimensional data drawn from a multivariate Normal (Gaussian) distribution.

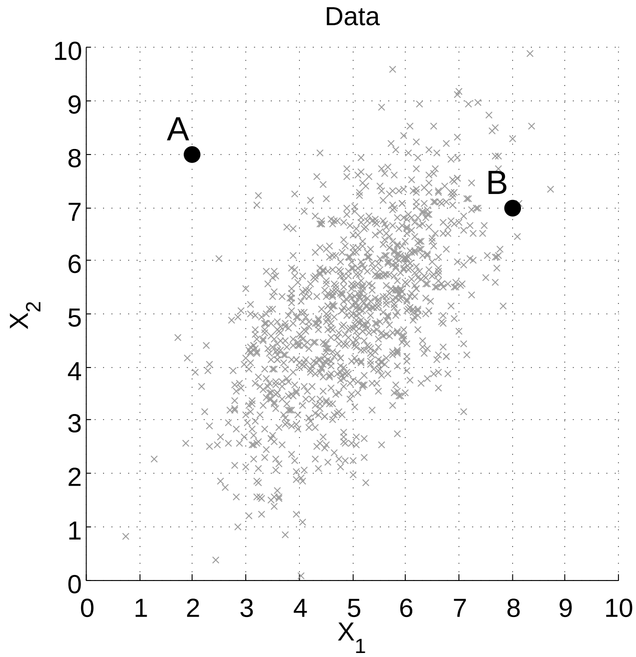


Figure 3: Two dimensional data drawn from a multivariate normal distribution.

10.1 The Multivariate Gaussian

1. **[Points: 2 pts]** What is the mean of this distribution? Estimate the answer visually and round to the nearest integer.

$$\mathbf{E}[X_1] = \mu_1 = 5\bigstar$$

$$\mathbf{E}[X_2] = \mu_2 = 5\bigstar$$

2. **[Points: 2 pts]** Would the off-diagonal covariance $\Sigma_{1,2} = \text{Cov}(X_1, X_2)$ be:

   (a) negative

   (b) positive ★

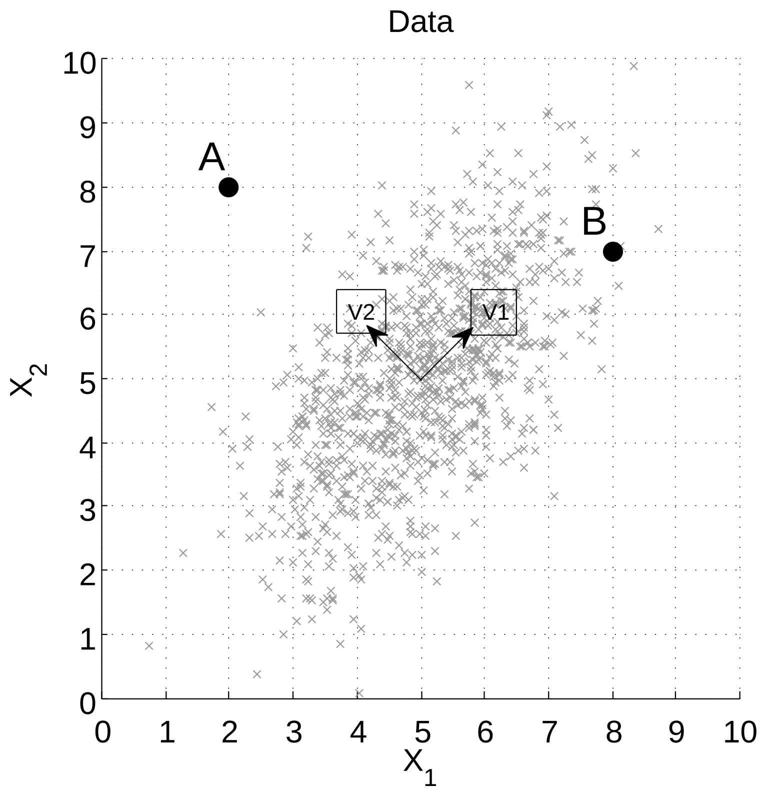
   (c) approximately zero

10.2 Principal Component Analysis

Define v_1 and v_2 as the directions of the first and second principal component, with $\|v_1\| = \|v_2\| = 1$. These directions define a change of basis

$$
\begin{aligned}
Z_1 &= (X - \mu) \cdot v_1 \\
Z_2 &= (X - \mu) \cdot v_2 \ .
\end{aligned}
$$

1. **[Points: 4 pts]** Sketch and label v_1 and v_2 on the following figure (a copy of Fig. 3). The arrows should originate from the mean of the distribution. You do not need to solve the SVD, instead visually estimate the directions.



★ **SOLUTION:** See above figure. Notice that both arrows are unit length.

2. **[Points: 2 pts]** The covariance $\text{Cov}(Z_1, Z_2)$, is (circle):
   (a) negative
   (b) positive
   (c) approximately zero ★

3. **[Points: 2 pts]** Which point (A or B) would have the higher reconstruction error after projecting onto the first principal component direction v_1? Circle one:

   Point A ★ Point B

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	$\frac{0+1}{1+2} = \frac{1}{3}$

12 Graphical Model Inference [8 Points]

Consider the following factor graph, simplified from the previous problem:

$$H \text{——} U \text{——} P$$

For this factor graph, suppose that we have learned the following potentials:

$\phi_1(H,U) =$

H	U	ϕ_1
T	T	3
T	F	1
F	T	2
F	F	0

$\phi_2(U,P) =$

U	P	ϕ_2
T	T	2
T	F	1
F	T	1
F	F	1

And, suppose that we observe, on a new evening, that $P = T$. Use variable elimination to determine $P(H \mid P = T)$. Please write your answer here:

$$\mathbf{P}(H = T \mid P = T) = \frac{7}{11}$$

$$\mathbf{P}(H = F \mid P = T) = \frac{4}{11}$$

And, please show your work in the following space:

★ **SOLUTION:** We first fix $P = T$ to derive the new factor:

$\phi_3(U) = \phi_2(U, P = T) =$

U	ϕ_3
T	2
F	1

Next we marginalize out U:

$$\begin{aligned}
\phi_4(H) &= \sum_{u \in \{T,F\}} \phi_1(H, U = u)\phi_3(U = u) \\
&= \phi_1(H, U = T)\phi_3(U = T) + \phi_1(H, U = F)\phi_3(U = F)
\end{aligned}$$

$=$

H	ϕ_4
T	$3 * 2 + 1 * 1 = 7$
F	$2 * 2 + 0 * 1 = 4$

Finally we normalize $\phi_4(H)$ to obtain the desired results:

$$\begin{aligned}
\mathbf{P}(H = T \mid P = T) &= 7/11 \\
\mathbf{P}(H = F \mid P = T) &= 4/11
\end{aligned}$$

13 Gibbs Sampling [16 Points]

In this problem you will use the factor graph in Fig. 4 along with the factors in Table 2. In addition you are given the normalizing constant Z defined as:

$$Z = \sum_{x_1=0}^{1} \sum_{x_2=0}^{1} \sum_{x_3=0}^{1} \sum_{x_4=0}^{1} \sum_{x_5=0}^{1} \sum_{x_6=0}^{1} f_1(x_1, x_2) f_2(x_1, x_3) f_3(x_1, x_4) f_4(x_2, x_5) f_5(x_3, x_4) f_6(x_4, x_6)$$

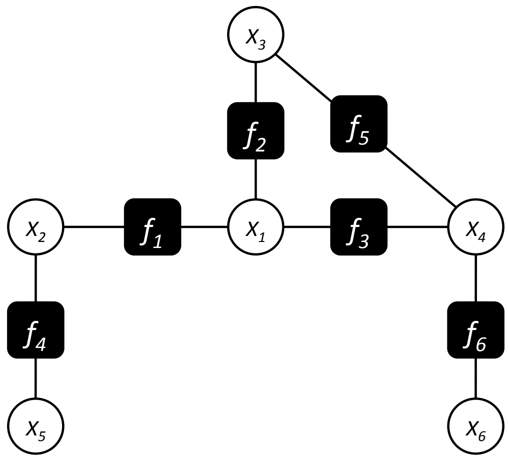


Figure 4: Simple factor graph with factors given in Table 2

f_1	$X_2 = 1$	$X_2 = 0$
$X_1 = 1$	a_1	b_1
$X_1 = 0$	c_1	d_1

f_2	$X_3 = 1$	$X_3 = 0$
$X_1 = 1$	a_2	b_2
$X_1 = 0$	c_2	d_2

f_3	$X_4 = 1$	$X_4 = 0$
$X_1 = 1$	a_3	b_3
$X_1 = 0$	c_3	d_3

f_4	$X_5 = 1$	$X_5 = 0$
$X_2 = 1$	a_4	b_4
$X_2 = 0$	c_4	d_4

f_5	$X_4 = 1$	$X_4 = 0$
$X_3 = 1$	a_5	b_5
$X_3 = 0$	c_5	d_5

f_6	$X_6 = 1$	$X_6 = 0$
$X_4 = 1$	a_6	b_6
$X_4 = 0$	c_6	d_6

Table 2: Factors for the factor graph in Fig. 4.

1. **[Points: 2 pts]** Circle the variables that are in the *Markov Blanket* of X_1:

   X_1 $\quad$ ★(X_2) $\quad$ ★(X_3) $\quad$ ★(X_4) $\quad$ X_5 $\quad$ X_6

2. **[Points: 2 pts]** What is the probability of the joint assignment:

   $\mathbf{P}(X_1 = 0, X_2 = 0, X_3 = 0, X_4 = 0, X_5 = 0, X_6 = 0) =$

   ★ **SOLUTION:** Don't forget the normalizing constant Z:

   $$\mathbf{P}(X_1 = 0, X_2 = 0, X_3 = 0, X_4 = 0, X_5 = 0, X_6 = 0) = \frac{1}{Z} d_1 d_2 d_3 d_4 d_5 d_6$$

3. **[Points: 4 pts]** In the Gibbs sampler, to draw a new value for X_1, we condition on its Markov Blanket. Suppose the current sample is $X_1 = 0, X_2 = 0, X_3 = 0, X_4 = 0, X_5 = 0, X_6 = 0$. What is:

   $\mathbf{P}(X_1 = 1 \mid \text{Markov Blanket of } X_1) =$

   ★ **SOLUTION:** The conditional equation is simply:

   $$\mathbf{P}(X_1 = 1 \mid X_2 = 0, X_3 = 0, X_4 = 0) = \frac{\mathbf{P}(X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 0)}{\mathbf{P}(X_1 = 0, X_2 = 0, X_3 = 0, X_4 = 0) + \mathbf{P}(X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 0)}$$

   Which is simply:

   $$\mathbf{P}(X_1 = 1 \mid X_2 = 0, X_3 = 0, X_4 = 0) = \frac{b_1 b_2 b_3}{d_1 d_2 d_3 + b_1 b_2 b_3}$$

4. **[Points: 2 pts]** (*Yes* or *No*) Do you need to know the normalizing constant for the joint distribution, Z, to be able to construct a Gibbs sampler?

   ★ **SOLUTION:** No. The Gibbs sampler only requires that you can compute the conditional of each variable given its Markov blanket.

5. **[Points: 6 pts]** After running the sampler for a while, the last few samples are as follows:

X_1	X_2	X_3	X_4	X_5	X_6
0	1	0	0	1	1
1	1	0	1	1	0
1	0	0	0	0	1
0	1	0	1	0	1

(a) Using the table, estimate $\mathbf{E}[X_6]$.

★ **SOLUTION:**

$$\mathbf{E}[X_6] = \frac{3}{4}$$

(b) Using the table, estimate $\mathbf{E}[X_1 X_5]$.

★ **SOLUTION:**

$$\mathbf{E}[X_1 X_5] = \frac{1}{4}$$

(c) Using the table, estimate $\mathbf{P}(X_1 = 1 \mid X_2 = 1)$.

★ **SOLUTION:**

$$\mathbf{P}(X_1 = 1 \mid X_2 = 1) = \frac{1}{3}$$

(d) Why might it be difficult to estimate $\mathbf{P}(X_1 = 1 \mid X_3 = 1)$ from the table?

★ **SOLUTION:** We do not have any samples for X_3. We would need to collect more samples to be able to estimate $\mathbf{P}(X_1 = 1 \mid X_3 = 1)$.

14 Extra Credit [22 Points]

You can gain extra credit in this course by answering any of the following questions.

14.1 Grow your Own Tree [14 Points]

You use your favorite decision tree algorithm to learn a decision tree for binary classification. Your tree has J leaves indexed $j = 1, \dots, J$. Leaf j contains n_j training examples, m_j of which are positive. However, instead of predicting a label, you would like to use this tree to predict the probability $P(Y = 1 \mid X)$ (where Y is the binary class and X are the input attributes). Therefore, you decide to have each leaf predict a real value $p_j \in [0, 1]$.

★ **SOLUTION:** We won't release the extra credit solutions yet. Since no one was able to get these questions fully right, they will be extra credit questions on Homework 5. Keep thinking about them! :)

1. **[Points: 2 pts]** What are the values p_j that yield the largest log likelihood? Show your work.

2. **[Points: 6 pts]** Now you decide to split the leaf j. You are considering splitting it into K new leaves indexed $k = 1, \dots, K$, each containing n'_k training examples, m'_k of which are positive (note that $\sum_k n'_k = n_j$, and $\sum_k m'_k = m_j$ since you are splitting the leaf j). What is the increase in log likelihood due to this split? Show your work and comment how it compares with the information gain.

3. **[Points: 6 pts]** The increase in log likelihood in the previous question can be used as a greedy criterion to grow your tree. However, in class you have learnt that maximum likelihood overfits. Therefore, you decide to incorporate recent results form learning theory and introduce the complexity penalty of the form

$$\lambda \sum_{j=1}^{J} \sqrt{n_j} \left| \log\left(\frac{p_j}{1-p_j} \right) \right| \ .$$

Now you optimize: negative log likelihood + penalty. What do you obtain as the optimal p_j? What do you use as the greedy splitting criterion? (You should be able to express the greedy criterion in a closed form using the optimal values for p_j before the split and optimal values for new leaves p'_k after the split.)

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.


END
DUALITY
GAP
SUPPORT
VECTOR
MACHINES
BAYESIANS
AGAINST
DISCRIMINATION