10-601 Machine Learning
Midterm Exam
Fall 2011

Tom Mitchell, Aarti Singh
Carnegie Mellon University

1. Personal information:
   - Name:
   - Andrew account:
   - E-mail address:
2. There should be **11** numbered pages in this exam.
3. This exam is open book, open notes. No computers or internet access is allowed.
4. You do not need a calculator.
5. If you need more room to answer a question, use the back of the page and clearly mark on the front of the page if we are to look at the back.
6. Work efficiently. Answer the easier questions first.
7. You have **80** minutes.
8. Good luck!

Question	Topic	Max. score	Score
1	Short questions	35	
2	MLE/MAP	15	
3	Bayes Nets	15	
4	EM	15	
5	Regression	20	
	Total	100	

1 Short Questions [35 pts]

Answer True/False in the following 8 questions. Explain your reasoning in 1 sentence.

1. **[3 pts]** Suppose you are given a dataset of cellular images from patients with and without cancer. If you are required to train a classifier that predicts the probability that the patient has cancer, you would prefer to use Decision trees over logistic regression.

   ★ **SOLUTION:** FALSE. Decision trees only provide a label estimate, whereas logistic regression provides the probability of a label (patient has cancer) for a given input (cellular image).

2. **[3 pts]** Suppose the dataset in the previous question had 900 cancer-free images and 100 images from cancer patients. If I train a classifier which achieves 85% accuracy on this dataset, it is it a good classifier.

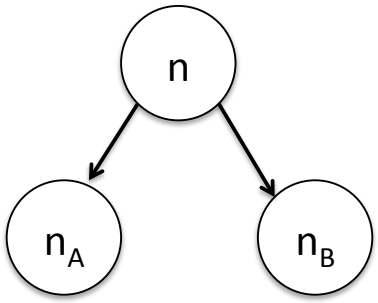
   ★ **SOLUTION:** FALSE. This is not a good accuracy on this dataset, since a classifier that outputs "cancer-free" for all input images will have better accuracy (90%).

3. **[3 pts]** A classifier that attains 100% accuracy on the training set and 70% accuracy on test set is better than a classifier that attains 70% accuracy on the training set and 75% accuracy on test set.

   ★ **SOLUTION:** FALSE. The second classifier has better test accuracy which reflects the true accuracy, whereas the first classifier is overfitting.

4. **[3 pts]** A football coach whispers a play number n to two players A and B independently. Due to crowd noise, each player imperfectly and independently draws a conclusion about what the play number was. A thinks he heard the number n_A, and B thinks he heard n_B. True or false: n_A and n_B are marginally dependent but conditionally independent given the true play number n.

   ★ **SOLUTION:** TRUE. Knowledge of n_A value tells us something about n_B therefore $P(n_A|n_B) \neq P(n_A)$ hence they are marginally dependent, but given n, n_A and n_B are determined independently. Also follows from following Bayes Net:



5. **[3 pts]** Assume m is the minimum number of training examples sufficient to guarantee that with probability $1-\delta$ a consistent learner using hypothesis space H will output a hypothesis with true error at worst ϵ. Then a second learner that uses hypothesis space H' will require $2m$ training examples (to make the same guarantee) if $|H'| = 2|H|$.

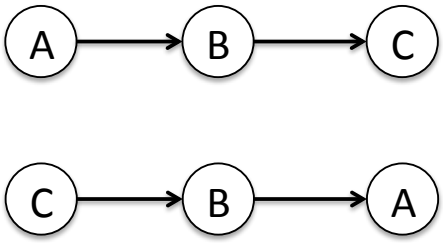
★ **SOLUTION:** FALSE. Minimum number of training examples sufficient to make an (ϵ, δ)-PAC guarantee depends logarithmically on hypothesis class size ($\ln |H|$) and not linearly.

6. **[3 pts]** If you train a linear regression estimator with only half the data, its bias is smaller.

F

★ **SOLUTION:** FALSE. Bias depends on the model you use (in this case linear regression) and not on the number of training data.

7. **[3 pts]** The following two Bayes nets encode the same set of conditional independence relations.



★ **SOLUTION:** TRUE. Both models encode that C and A are conditionally independent given B. Also

$$P(A)P(B|A)P(C|B) = \frac{P(A,B)P(B,C)}{P(B)} = P(C)P(B|C)P(A|B)$$

8. **[3 pts]** A, B and C are three Boolean random variables. The following equality holds without any assumptions on the joint distribution $P(A, B, C)$

$$P(A|B) = P(A|B, C = 0)P(C = 0) + P(A|B, C = 1)P(C = 1).$$

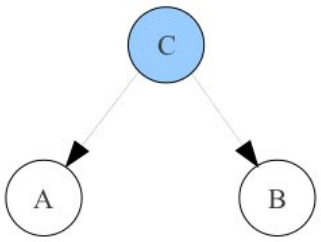
★ **SOLUTION:** TRUE. Since C is a Boolean random variable, we have

$$\begin{aligned} P(A|B) &= P(A, C = 0|B) + P(A, C = 1|B) \\ &= P(A|B, C = 0)P(C = 0) + P(A|B, C = 1)P(C = 1) \end{aligned}$$

where last step follows from definition of conditional probability.

The following three short questions are not True/False questions. Please provide explanations for your answers.

9. **[3 pts]** The Bayes net below implies that A is conditionally independent of B given C ($A \perp\!\!\!\perp B|C$). Prove this, based on its factorization of the joint distribution, and on the definition of conditional independence.



★ **SOLUTION:** Using factorization of joint distribution

$$P(A, B, C) = P(C)P(A|C)P(B|C)$$

and using definition of conditional independence

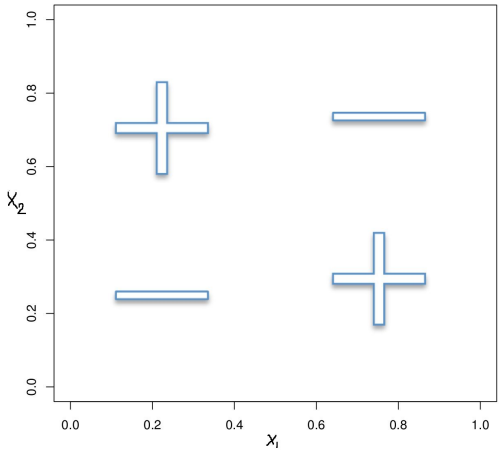
$$P(A, B, C) = P(C)P(A, B|C)$$

Therefore, we have:

$$P(A, B|C) = P(A|C)P(B|C)$$

i.e. A is conditionally independent of B given C ($A \perp\!\!\!\perp B|C$).

10. [**3 pts**] Which of the following classifiers can perfectly classify the following data:



(a) Decision Tree

(b) Logistic Regression

(c) Gaussian Naive Bayes

★ **SOLUTION:** Decision Tree only. Decision trees of depth 2 which first splits on X_1 and then on X_2 wil perfectly classify it. Logistic regression leads to linear decision boundaries, hence cannot classify this data perfectly. Due to conditional independence requirement, it is not possible to fit a Gaussian that peaks at the labels of only one class and has no covariance between features, so Gaussian Naive Bayes cannot classify this data perfectly.

11. [**5 pts**] Boolean random variables A and B have the joint distribution specified in the table below.

A	B	$P(A, B)$
0	0	0.32
0	1	0.48
1	0	0.08
1	1	0.12

Given the above table, please compute the following five quantities:

★ **SOLUTION:** $P(A = 0) = P(A = 0, B = 0) + P(A = 0, B = 1) = 0.32 + 0.48 = 0.8$

$P(A = 1) = 1 - P(A = 0) = 0.2$

$P(B = 1) = P(B = 1, A = 0) + P(B = 1, A = 1) = 0.48 + 0.12 = 0.6$

$P(B = 0) = 1 - P(B = 1) = 0.4$

$P(A = 1|B = 0) = P(A = 1, B = 0)/P(B = 0) = 0.08/0.4 = 0.2$

Are A and B independent? Justify your answer.

★ **SOLUTION:** YES. Using the calculations above,

$$P(A = 0)P(B = 0) = 0.8 * 0.4 = 0.32 = P(A = 0, B = 0)$$
$$P(A = 0)P(B = 1) = 0.8 * 0.6 = 0.48 = P(A = 0, B = 1)$$
$$P(A = 1)P(B = 0) = 0.2 * 0.4 = 0.08 = P(A = 1, B = 0)$$
$$P(A = 1)P(B = 1) = 0.2 * 0.6 = 0.12 = P(A = 1, B = 1)$$

2 MLE/MAP Estimation [15 pts]

In this question you will estimate the probability of a coin landing heads using MLE and MAP estimates.

Suppose you have a coin whose probability of landing heads is $p = 0.5$, that is, it is a fair coin. However, you do not know p and would like to form an estimator $\hat{\theta}$ for the probability of landing heads p. In class, we derived an estimator that assumed p can take on any value in the interval $[0, 1]$. In this question, you will derive an estimator that assumes p can take on only two possible values: 0.3 or 0.6.

Note: $P_{\hat{\theta}}[\text{heads}] = \hat{\theta}$.

Hint: All the calculations involved here are simple. You do not require a calculator.

1. **[5 pts]** You flip the coin 3 times and note that it landed 2 times on tails and 1 time on heads. Find the maximum likelihood estimate $\hat{\theta}$ of p over the set of possible values $\{0.3, 0.6\}$.

   Solution:

   $$\begin{aligned}\hat{\theta} &= \text{argmax}_{\theta \in \{0.3, 0.6\}}\ P_\theta[D] \\ &= \text{argmax}_{\theta \in \{0.3, 0.6\}}\ P_\theta[\text{heads}] P_\theta[\text{tails}]^2 \\ &= \text{argmax}_{\theta \in \{0.3, 0.6\}}\ \theta(1-\theta)^2\end{aligned}$$

   We observe that

   $$\frac{P_{\theta=0.3}[D]}{P_{\theta=0.6}D]} = \frac{0.3 * 0.7^2}{0.6 * 0.4^2} = \frac{0.49}{0.32} > 1$$

   which implies that $\hat{\theta} = 0.3$.

2. **[4 pts]** Suppose that you have the following prior on the parameter p:

   $$P[p = 0.3] = 0.3 \quad \text{and} \quad P[p = 0.6] = 0.7.$$

   Again, you flip the coin 3 times and note that it landed 2 times on tails and 1 time on heads. Find the MAP estimate $\hat{\theta}$ of p over the set $\{0.3, 0.6\}$, using this prior.

   Solution:

   $$\hat{\theta} = \text{argmax}_{\theta \in \{0.3, 0.6\}}\ P_\theta[D] P[\theta]$$

   We observe that

   $$\frac{P_{\theta=0.3}[D] P[\theta = 0.3]}{P_{\theta=0.6}[D] P[\theta = 0.6]} = \frac{0.3 * 0.7^2 * 0.3}{0.6 * 0.4^2 * 0.7} = \frac{0.21}{0.32} < 1$$

   which implies that $\hat{\theta}_{\text{MAP}} = 0.6$.

3. **[3 pts]** Suppose that the number of times you flip the coin tends to infinity. What would be the maximum likelihood estimate $\hat{\theta}$ of p over the set $\{0.3, 0.6\}$ in that case? Justify your answer.

   Solution:
   With the number of flips tending to infinity, proportion of heads to the total number of flips tends to 0.5. The MLE would be 0.6 as this is closer to 0.5.

4. **[3 pts]** Suppose that the number of times you flip the coin tends to infinity. What would be the MAP estimate $\hat{\theta}$ of p over the set $\{0.3, 0.6\}$, using the prior defined in part 2 of this question? Justify your answer.

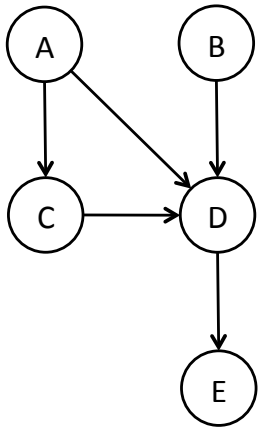
   Solution:
   With the number of flips tending to infinity, the effect of the prior becomes negligible. Therefore, the MAP estimate will be the same as the MLE.

3 Bayes Nets [15 pts]

1. (a) **[3 pts]** Please draw a Bayes net which represents the following joint distribution:

$$P(A, B, C, D, E) = P(A)P(B)P(C|A)P(D|A, B, C)P(E|D)$$

Solution:



(b) **[2 pts]** For the graph that you drew above, assume each variable can take on the values 1, 2 or 3. Also assume that you are given values for the probabilities $P(D = 1)$, $P(D = 2)$ and $P(D = 3)$. Please specify the smallest set of Bayes net parameters you would need in order to calculate $P(E = 1)$. *Solution:*
We can write:

$$\begin{aligned} P(E = 1) =& P(E = 1|D = 1)P(D = 1) + P(E = 1|D = 2)P(D = 2) + \dots \\ & P(E = 1|D = 3)P(D = 3) \end{aligned}$$

Thus we need three parameters: $P(E = 1|D = 1)$, $P(E = 1|D = 2)$ and $P(E = 1|D = 3)$.

2. **[4 pts]** Please draw a single Bayes net which encodes all the following conditional independence assumptions over the variables A, B, C and D:

   (a) A is independent of D given B and C

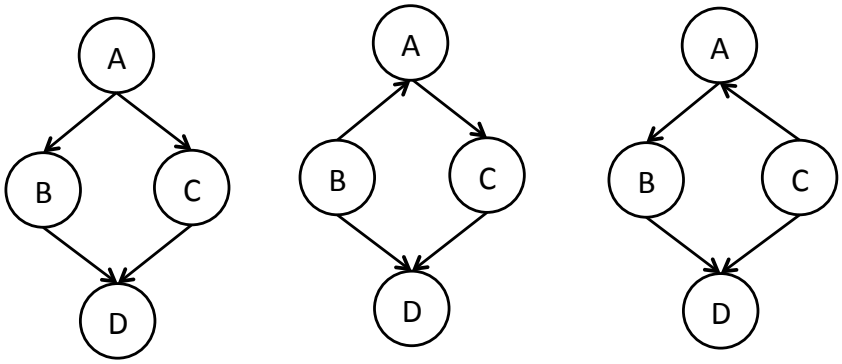
   (b) A is *not* independent of D given only B

   (c) A is *not* independent of D given only C

(d) B is independent of C given only A

(e) B is *not* independent of C given A and D

Solution: Any of the following satisfy the above:



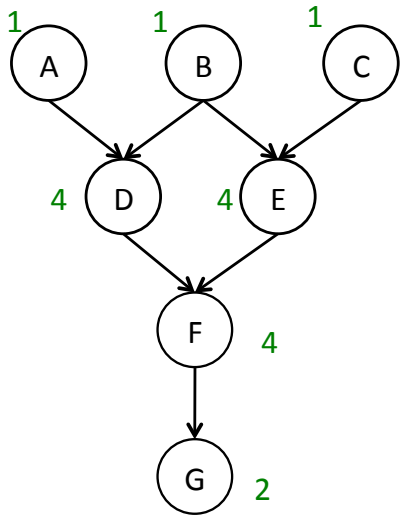
3. [**4 pts**] Consider the graph drawn below. Assume that each variable can only take on values *true* and *false*.

   (a) How many parameters are necessary to specify the joint distribution $P(A, B, C, D, E, F, G)$ for this Bayes net? You may answer by writing the number of parameters directly next to each graph node.

   Solution: See below for the number of parameters needed for each node. Total is 17.

   (b) Please give the **minimum** number of Bayes net parameters required to fully specify the distribution $P(G|A, B, C, D, E, F)$. Briefly justify your answer.

   Solution: Note that the Markov blanket for G consists only of F. Thus, $P(G|A, B, C, D, E, F) = P(G|F)$ and only two parameters are need to specify this distribution.



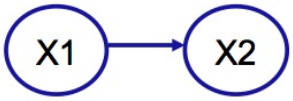
4. [**2 pts**] Given the graph provided above, please state if the following are **true** or **false**.

   (a) E is conditionally independent of G given F. *Solution:* True.

   (b) A is conditionally independent of C given B and G. *Solution:* False.

4 EM [15 pts]

In this question you will apply EM to train the following simple Bayes net:



using the following data set, for which $X2$ is unobserved in training example 4.

Example	$X1$	$X2$
1.	0	1
2.	0	0
3.	1	0
4.	1	?
5.	0	1

The EM process has run for several iterations. At this point the parameter estimates are:

$$\hat{\theta}_{X1=1} = \hat{P}(X1 = 1) = 0.4$$
$$\hat{\theta}_{X2=1|X1=1} = \hat{P}(X2 = 1|X1 = 1) = 0.4$$
$$\hat{\theta}_{X2=1|X1=0} = \hat{P}(X2 = 1|X1 = 0) = 0.66$$

1. **[2 pts]** What is calculated in the next E step?

   Answer: The expected value of $X2$ for example 4: $P(X2 = 1|X1 = 1; \theta)$

2. **[5 pts]** What precisely is the result of the next E step? Show your work.

$$\hat{P}(X2 = 1|X1 = 1) = \hat{\theta}_{X2=1|X1=1} = 0.4$$

3. **[3 pts]** What is calculated in the next M step?

   New estimates for $\hat{\theta}_{X1=1}$, $\hat{\theta}_{X2=1|X1=0}$ (which do not change), and $\hat{\theta}_{X2=1|X1=1}$

4. **[5 pts]** What precisely is the result of the next M step? Show your work.

$$\hat{\theta}_{X1=1} = \frac{2}{5} = 0.4$$
$$\hat{\theta}_{X2=1|X1=0} = \frac{2}{3} = 0.66$$
$$\hat{\theta}_{X2=1|X1=1} = \frac{0.4}{2} = 0.2$$

5 Bias and Variance in Linear Regression [20 pts]

In this question, we will explore bias and variance in linear regression. Assume that a total of N data points of the form (x_i, y_i) are generated from the following (true) model:

$$x_i \sim Unif(0,1), \quad y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0,1), \quad f(x) = x$$

We assume $x_i \perp \epsilon_j \; \forall i, j$ and $\epsilon_i \perp \epsilon_j \; \forall i \neq j$ (note $a \perp b$ means a and b are independent).

You may find the following pieces of information useful when solving this problem:

- $bias^2 = \int_x (E_D[h_D(x)] - f(x))^2 p(x) dx$
- $variance = \int_x E_D[(h_D(x) - E_D[h_D(x)])^2] p(x) dx$
- $\hat{\mu} \sim N(\mu, \frac{1}{N})$ if $\hat{\mu}$ is the MLE estimator with N data points
- If $x \sim Unif(0,1)$, then $\int_0^1 p(x)dx \; = 1$, and therefore $p(x) = 1$.

We begin by examining the case where we are not aware that y depends on x. Instead, our (incorrect) model is that $f(x)$ has some constant value $f(x) = \mu$, and therefore

$$x_i \sim Unif(0,1), \quad y_i \sim N(\mu, 1) \text{ with } x_i \perp y_i.$$

We use the MLE estimator for μ. That is, we let $\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} y_i$. The prediction of our trivial regression model for the value of y_i is $\hat{\mu}$, regardless of the value of x_i.

1. **[2 pts]** What is the value for $E_D[h_D(x)]$ in this case? Here E_D refers to the expected value over different training data sets of size N, and $h_D(x)$ is the predictor learned from a specific data set D.

   ★ **SOLUTION:** $E_D[h_D(x)] = E_D[\hat{\mu}] = \frac{1}{2}$

2. **[3 pts]** What is the bias of this trivial regression model?

   ★ **SOLUTION:** $Bias^2 = \int_0^1 (\frac{1}{2} - x)^2 (1) \; dx = -\frac{1}{3}(\frac{1}{2} - x)^3 |_0^1 = \frac{1}{12}$.
   The bias is thus $\sqrt{\frac{1}{12}}$.

3. **[2 pts]** What is the variance of this trivial regression model?

★ **SOLUTION:** The variance is the variance of the MLE estimator. By the third bullet, this is $\frac{1}{N}$.

4. [**1 pts**] What is the unavoidable error in this learning setting?

   ★ **SOLUTION:** The unavoidable error is introduced by ϵ_i, and is 1 by assumption.

5. [**2 pts**] How do each of bias, variance, and unavoidable error change as $N \to \infty$?

   ★ **SOLUTION:** The unavoidable error and bias do not change. The variance goes to 0 as $N \to \infty$.

Now assume we notice that y in fact depends on x. Therefore, we change to a linear regression model (with zero intercept), which assumes the data are generated as follows:

$$x_i \sim Unif(0,1), \quad y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0,1), \quad f(x) = ax$$

We also assume (as in the true model) that $x_i \perp \epsilon_j \ \forall i, j$ and $\epsilon_i \perp \epsilon_j \ \forall i \neq j$.

6. **[3 pts]** We choose our estimator $\hat{a}$ for a to minimize the squared sum of errors. That is, we choose $\hat{a}$ such that

$$\hat{a} = argmin_a \ \frac{1}{2} \sum_{i=1}^{N} (y_i - ax_i)^2$$

Derive the closed form expression for $\hat{a}$. Once we have chosen the value of $\hat{a}$, we now have a regression model that predicts $y_i = \hat{a}x_i$.

★ **SOLUTION:** Let $f(a) = \frac{1}{2} \sum_{i=1}^{N} (y_i - ax_i)^2$. Then,

$\frac{\partial(f)}{\partial a} = \sum_{i=1}^{N} -x_i(y_i - ax_i)$

Setting the derivative to 0, we obtain:

$\sum_{i=1}^{N} -x_i(y_i - ax_i) = 0$

$\implies \sum_{i=1}^{N} x_i y_i = \sum_{i=1}^{N} a x_i^2$

$$\implies \hat{a} = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}.$$

7. **[2 pts]** What is the bias of this linear regression model?

★ **SOLUTION:** The bias of the regression model is 0.

8. **[2 pts]** As $N \to \infty$, what is the variance of this linear regression model?

★ **SOLUTION:** The variance of the linear regression models goes to 0 as $N \to \infty$.

9. **[1 pts]** What is the unavoidable error in this learning setting?

★ **SOLUTION:** The unavoidable error is still introduced by ϵ_i, and is 1 by assumption.

10. [**2 pts**] In the figure below, draw the two learned regression models if we have an infinite number of data points.

★ **SOLUTION:** Model 1 (the trivial model) is the horizontal line. Model 2 (the linear regression model) is the diagonal line.