10-701 Midterm Exam, Spring 2011

1. Personal info:
   - Name:
   - Andrew account:
   - E-mail address:
2. There are 14 numbered pages in this exam (including this cover sheet).
3. You can use any material you brought: any book, notes, and print outs. You cannot use materials brought by other students.
4. No computers, PDAs, phones or Internet access.
5. If you need more room to answer a question, use the back of the preceding page.
6. Work efficiently. Consider answering all of the easier questions first.
7. There is one *optional extra credit question*, which will *not* affect the grading curve. It will be used to bump your grade up, without affecting anyone else's grade.
8. You have 80 minutes, the test has 100 points. Good luck!

Question	Topic	Max. score	Score
1	Short Questions	20	
2	Bayes Nets	23	
3	Decision Surfaces and Training Rules	12	
4	Linear Regression	20	
5	Conditional Independence Violation	25	
6	[Extra Credit] Violated Assumptions	6	

1 [20 Points] Short Questions

1.1 True or False (Grading: Carl Doersch)

Answer each of the following True of False. If True, give a short justification. If False, a counter-example or convincing one-sentence explanation.

1. [2 pts] If we train a Naive Bayes classifier using infinite training data that satisfies all of its modeling assumptions (e.g., conditional independence), then it will achieve zero *training error* over these training examples.

   ★ **SOLUTION:** **False**. There will still be unavoidable error. In Naive Bayes, Y is probabilistic, so it is often impossible to predict Y even if the model's estimate of $P(Y)$ is perfect. Furthermore, Naive Bayes is linear, and so it can't necessarily even estimate $P(Y)$ perfectly: for example, in the distribution $Y = 1 \Leftrightarrow X_1 \text{XOR} X_2$.

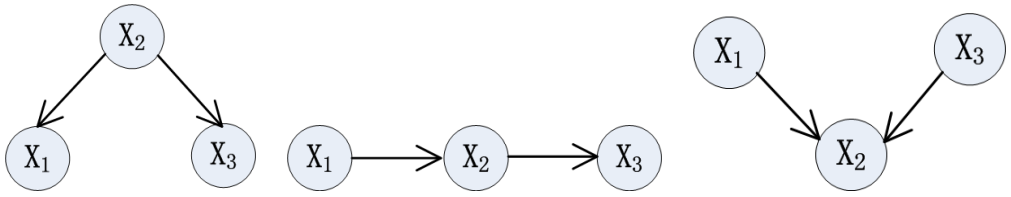
2. [2 pts] If we train a Naive Bayes classifier using infinite training data that satisfies all of its modeling assumptions (e.g., conditional independence), then it will achieve zero *true error* over test examples drawn from this same distribution.

   ★ **SOLUTION:** **False**, for the same reasons as above.

3. [2 pts] Every Bayes Net defined over 10 variables $\langle X_1, X_2, \ldots X_{10} \rangle$ tells how to factor the joint probability distribution $P(X_1, X_2, \ldots X_{10})$ into the product of exactly 10 terms.

   ★ **SOLUTION:** **True**, by the definition of Bayes Net.

   Consider the three Bayes Nets shown below:



A B C

4. [3 pts] True or false: Every joint distribution $P(X_1, X_2, X_3)$ that can be defined by adding Conditional Probability Distributions (CPD) to Bayes Net graph A can also be expressed by appropriate CPD's for Bayes Net graph B.

★ **SOLUTION:** **True**. If a distribution can be represented in graph A, it will factorize as $P(X_2)P(X_1|X_2)P(X_3|X_2)$. Using Bayes rule, this becomes $P(X_2)P(X_3|X_2)P(X_2|X_1)P(X_1)/P(X_2) = P(X_3|X_2)P(X_2|X_1)P(X_1)$.

5. [3 pts] True or false: Every joint distribution $P(X_1, X_2, X_3)$ that can be defined by adding Conditional Probability Distributions to Bayes Net graph A can also be expressed by appropriate CPD's for Bayes Net graph C.

★ **SOLUTION:** **False**. A can represent distributions where X_1 can depend on X_3 given no information about X_2, whereas graph C cannot.

1.2 Quick questions (Grading: Yi Zhang)

Answer each of the following in one or two sentences, in the space provided.

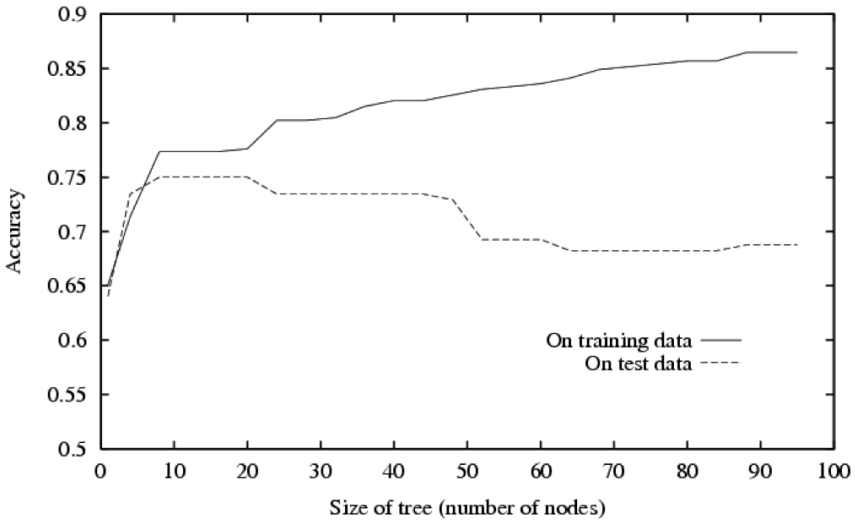
1. [2 pts] Prove that $P(X_1|X_2)P(X_2) = P(X_2|X_1)P(X_1)$. (*Hint:* This is a two-line proof.)

★ **SOLUTION:** $P(X_1|X_2)P(X_2) = P(X_1, X_2) = P(X_2|X_1)P(X_1)$

2. [3 pts] Consider a decision tree learner applied to data where each example is described by 10 boolean variables $\langle X_1, X_2, \ldots X_{10}\rangle$. What is the VC dimension of the hypothesis space used by this decision tree learner?

★ **SOLUTION:** The VC dimension is 2^{10}, because we can shatter 2^{10} examples using a tree with 2^{10} leaf nodes, and we cannot shatter $2^{10} + 1$ examples (since in that case we must have duplicated examples and they can be assigned with conflicting labels).

3. [3 pts] Consider the plot below showing training and test set accuracy for decision trees of different sizes, using the same set of training data to train each tree. Describe in one sentence how the training data curve (solid line) will change if the *number of training examples* approaches infinity. In a second sentence, describe what will happen to the test data curve under the same condition.

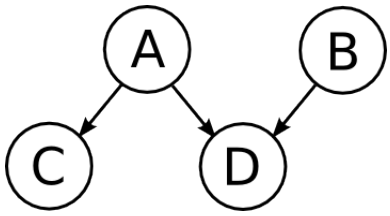


★ **SOLUTION:** The new training accuracy curve should be below the original training curve (since it's impossible for the trees to overfit infinite training data); the new testing accuracy curve should be above the original testing curve and become identical to the new training curve (since trees learned from infinite training data should perform well on testing data and do not overfit at all).

2 [23 Points] Bayes Nets (Grading: Carl Doersch)

2.1 [17 pts] Inference

In the following graphical model, A, B, C, and D are binary random variables.



1. [2 pts] How many parameters are needed to define the Conditional Probability Distributions (CPD's) for this Bayes Net?

   ★ **SOLUTION:** 8: 1 for A, 1 for B, 2 for C, and 4 for D.

2. [2 pts] Write an expression for the probability $P(A = 1, B = 1, C = 1, D = 1)$ in terms of the Bayes Net CPD parameters. Use notation like $P(C = 1|A = 0)$ to denote specific parameters in the CPD's.

   ★ **SOLUTION:**

   $$P(A = 1)P(B = 1)P(C = 1|A = 1)P(D = 1|A = 1, B = 1)$$

3. [3 pts] Write an expression for $P(A = 0|B = 1, C = 1, D = 1)$ in terms of the Bayes Net Conditional Probability Distribution (CPD) parameters.

   ★ **SOLUTION:**

   $$\frac{P(A = 0)P(B = 1)P(C = 1|A = 0)P(D = 1|A = 0, B = 1)}{P(A = 0)P(B = 1)P(C = 1|A = 0)P(D = 1|A = 0, B = 1)+ \\ P(A = 1)P(B = 1)P(C = 1|A = 1)P(D = 1|A = 1, B = 1)}$$

4. [2 pts] True or False (give brief justification): C is conditionally independent of B given D.

★ **SOLUTION:** **False**. There is one path from C to B, and this path isn't blocked at either node.

5. [2 pts] True or False (give brief justification): C is conditionally independent of B given A.

★ **SOLUTION:** **True**. The path is now blocked at both A and D.

Suppose we use EM to train the above Bayes Net from the partially labeled data given below, first initializing all Bayes net parameters to 0.5.

A	B	C	D
1	0	1	0
1	?	0	1
1	1	0	?
0	?	0	?
0	1	0	?

6. [2 pts] How many distinct quantities will be updated during the first M step?

★ **SOLUTION:** **5** or **8**, depending on your interpretation. In the M step we update the values of all parameters, and from part 1 there were 8 parameters. However, only 5 of them will actually be changed if your algorithm's initialization is clever.

7. [2 pts] How many distinct quantities will be estimated during the first E step?

★ **SOLUTION:** **5**. Every unknown value must be estimated.

8. [2 pts] When EM converges, what will be the final estimate for $P(C = 0|A = 1)$? [*Hint:* You do not need a calculator.]

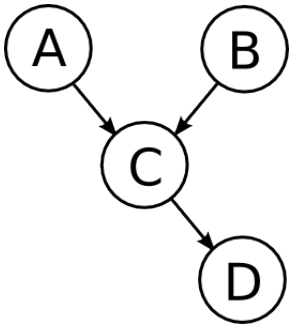
★ **SOLUTION:** **2/3**: the fraction of times when C=0 out of all examples where A=1.

2.2 [6 pts] Constructing a Bayes net

Draw a Bayes net over the random variables $\{A, B, C, D\}$ where the following conditional independence assumptions hold. Here, $X \perp Y | Z$ means X is conditionally independent of Y given Z, and $X \not\perp Y | Z$ means X and Y are not conditionally independent given Z, and $\emptyset$ stands for the empty set.

- $A \perp B | \emptyset$
- $A \not\perp D | B$
- $A \perp D | C$
- $A \not\perp C | \emptyset$
- $B \not\perp C | \emptyset$
- $A \not\perp B | D$
- $B \perp D | A, C$

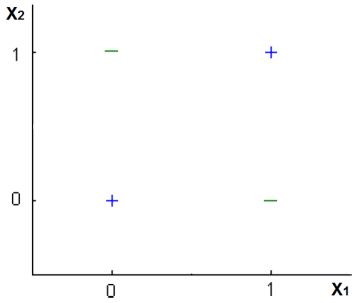
★ **SOLUTION:**



3 [12 Points] Decision Surfaces and Training Rules (Grading: Yi Zhang)

Consider a classification problem with two boolean variables $X_1, X_2 \in \{0, 1\}$ and label $Y \in \{0, 1\}$. In Figure 1 we show two positive ("+") and two negative ("-") examples.

Figure 1: Two positive examples and two negative examples.



Question [2 pts]: Draw (or just simply describe) a decision tree that can perfectly classify the four examples in Figure 1.

★ **SOLUTION:** Split using one variable (e.g., X_1) and then split using the other variable (e.g., X_2). Label each leaf node according to the assigned training example.

Question [3 pts]: In the class we learned the training rule to grow a decision tree: we start from a single root node and iteratively split each node using the "best" attribute selected by maximizing the information gain of the split. We will stop splitting a node if: 1) examples in the node are already pure; or 2) we cannot find any single attribute that gives a split with *positive* information gain. If we apply this training rule to the examples in Figure 1, will we get a decision tree that perfectly classifies the examples? Briefly explain what will happen.

★ **SOLUTION:** We will stop at a single root node and cannot grow the tree any more. This is because, at the root node, splitting on any single variable has zero information gain.

Question [5 pts]: Suppose we learn a Naive Bayes classifier from the examples in Figure 1, using MLE (maximum likelihood estimation) as the training rule. Write down all the parameters and their estimated values (note: both $P(Y)$ and $P(X_i|Y)$ should be Bernoulli distributions). Also, does this learned Naive Bayes perfectly classify the four examples?

★ **SOLUTION:** $P(Y = 1) = 0.5$ $(= P(Y = 0))$
$P(X_1 = 1|Y = 0) = P(X_1 = 1|Y = 1) = 0.5$ $(= P(X_1 = 0|Y = 0) = P(X_1 = 0|Y = 1))$
$P(X_2 = 1|Y = 0) = P(X_2 = 1|Y = 1) = 0.5$ $(= P(X_2 = 0|Y = 0) = P(X_2 = 0|Y = 1))$
This is a very poor classifier since for any X_1, X_2 it will predict $P(Y = 1|X_1, X_2) = P(Y = 0|X_1, X_2) = 0.5$. Naturally, it cannot perfectly classify the examples in the figure.

Question [2 pts]: Is there any logistic regression classifier using X_1 and X_2 that can perfectly classify the examples in Figure 1? Why?

★ **SOLUTION:** No, logistic regression only forms linear decision surface, but the examples in the figure are not linearly separable.

4 [20 Points] Linear Regression (Grading: Xi Chen)

Consider a simple linear regression model in which y is the sum of a deterministic linear function of x, plus random noise ϵ.

$$y = wx + \epsilon$$

where x is the real-valued input; y is the real-valued output; and w is a single real-valued parameter to be learned. Here ϵ is a real-valued random variable that represents noise, and that follows a Gaussian distribution with mean 0 and standard deviation σ; that is, $\epsilon \sim N(0, \sigma)$

(a) [3pts] Note that y is a random variable because it is the sum of a deterministic function of x, plus the random variable ϵ. Write down an expression for the probability distribution governing y, in terms of $N(), \sigma, w$ and x.

★ **SOLUTION:** y follows a Gaussian distribution with the mean wx and the standard deviation σ:

$$p(y|w, x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y - wx)^2}{2\sigma^2}\right\}$$

(b) [3 pts] You are given n i.i.d. training examples $\{(x^1, y^1), (x^2, y^2), \ldots, (x^n, y^n)\}$ to train this model. Let $\mathcal{Y} = (y^1, \ldots, y^n)$ and $\mathcal{X} = (x^1, \ldots, x^n)$, write an expression for the conditional data likelihood: $p(\mathcal{Y}|\mathcal{X}, w)$.

★ **SOLUTION:**

$$\begin{aligned} p(\mathcal{Y}|\mathcal{X}, w) &= \prod_{i=1}^{n} p(y^i|x^i, w) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \prod_{i=1}^{n} \exp\left\{-\frac{(y^i - wx^i)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{\sum_{i=1}^{n}(y^i - wx^i)^2}{2\sigma^2}\right\} \end{aligned}$$

(c) [9 pts] Here you will derive the expression for obtaining a MAP estimate of w from the training data. Assume a Gaussian prior over w with mean 0 and standard deviation τ (i.e. $w \sim N(0, \tau)$). Show that finding the MAP estimate w^* is equivalent to solving the following optimization problem:

$$w^* = \text{argmin}_w \frac{1}{2} \sum_{i=1}^{n} (y^i - wx^i)^2 + \frac{\lambda}{2} w^2;$$

Also express the regularization parameter λ in terms of σ and τ.

★ SOLUTION:

$$\begin{aligned} p(w|\mathcal{Y}, \mathcal{X}) &\propto p(\mathcal{Y}|\mathcal{X}, w) p(w|\mathcal{X}) \\ &\propto \exp\left\{ -\frac{\sum_{i=1}^{n} (y^i - wx^i)^2}{2\sigma^2} \right\} \exp\left\{ -\frac{w^2}{2\tau^2} \right\} \end{aligned}$$

$$\begin{aligned} w^* &= \text{argmax}_w \ln p(w|\mathcal{Y}, \mathcal{X}) \\ &= \text{argmax}_w - \frac{\sum_{i=1}^{n} (y^i - wx^i)^2}{2\sigma^2} - \frac{w^2}{2\tau^2} \\ &= \text{argmin}_w \frac{\sum_{i=1}^{n} (y^i - wx^i)^2}{2\sigma^2} + \frac{w^2}{2\tau^2} \\ &= \text{argmin}_w \frac{1}{2} \sum_{i=1}^{n} (y^i - wx^i)^2 + \frac{\sigma^2}{2\tau^2} w^2 \end{aligned}$$

We can see that $\lambda = \frac{\sigma^2}{\tau^2}$.

(d) [5pts] Above we assumed a zero-mean prior for w, which resulted in the usual $\frac{\lambda}{2} w^2$ regularization term for linear regression. Sometimes we may have prior knowledge that suggests w has some value other than zero. Write down the revised objective function that would be derived if we assume a Gaussian prior on w with mean μ instead of zero (i.e., if the prior is $w \sim N(\mu, \tau)$).

★ SOLUTION:

$$\begin{aligned} p(w|\mathcal{Y}, \mathcal{X}) &\propto p(\mathcal{Y}|\mathcal{X}, w) p(w|\mathcal{X}) \\ &\propto \exp\left\{ -\frac{\sum_{i=1}^{n} (y^i - wx^i)^2}{2\sigma^2} \right\} \exp\left\{ -\frac{(w-\mu)^2}{2\tau^2} \right\} \end{aligned}$$

$$\begin{aligned} w^* &= \text{argmax}_w \ln p(w|\mathcal{Y}, \mathcal{X}) \\ &= \text{argmax}_w - \frac{\sum_{i=1}^{n} (y^i - wx^i)^2}{2\sigma^2} - \frac{(w-\mu)^2}{2\tau^2} \\ &= \text{argmin}_w \frac{\sum_{i=1}^{n} (y^i - wx^i)^2}{2\sigma^2} + \frac{(w-\mu)^2}{2\tau^2} \\ &= \text{argmin}_w \frac{1}{2} \sum_{i=1}^{n} (y^i - wx^i)^2 + \frac{\sigma^2}{2\tau^2} (w-\mu)^2 \end{aligned}$$

5 [25 Points] Conditional Independence Violation (Grading: Yi Zhang)

5.1 Naive Bayes without Conditional Independence Violation

Table 1: $P(Y)$

$Y=0$	$Y=1$
0.8	0.2

Table 2: $P(X_1|Y)$

	$X_1=0$	$X_1=1$
$Y=0$	0.7	0.3
$Y=1$	0.3	0.7

Consider a binary classification problem with variable $X_1 \in \{0,1\}$ and label $Y \in \{0,1\}$. The true generative distribution $P(X_1, Y) = P(Y)P(X_1|Y)$ is shown as Table 1 and Table 2.

Question [4 pts]: Now suppose we have trained a Naive Bayes classifier, using *infinite* training data generated according to Table 1 and Table 2. In Table 3, please write down the predictions from the trained Naive Bayes for different configurations of X_1. Note that $\hat{Y}(X_1)$ in the table is the decision about the value of Y given X_1. For decision terms in the table, write down either $\hat{Y}=0$ or $\hat{Y}=1$; for probability terms in the table, write down the actual values (and the calculation process if you prefer, e.g., $0.8 * 0.7 = 0.56$).

Table 3: Predictions from the trained Naive Bayes

	$\hat{P}(X_1, Y=0)$	$\hat{P}(X_1, Y=1)$	$\hat{Y}(X_1)$
$X_1=0$	$0.8 \times 0.7 = 0.56$	$0.2 \times 0.3 = 0.06$	$\hat{Y}=0$
$X_1=1$	$0.8 \times 0.3 = 0.24$	$0.2 \times 0.7 = 0.14$	$\hat{Y}=0$

★ **SOLUTION:** The naive Bayes model learned from infinite data will have $\hat{P}(Y)$ and $\hat{P}(X_1|Y)$ estimated exactly as Table 1 and Table 2. The resulting predictions are shown in Table 3.

Question [3 pts]: What is the expected error rate of this Naive Bayes classifier on testing examples that are generated according to Table 1 and Table 2? In other words, $P(\hat{Y}(X_1) \neq Y)$ when (X_1, Y) is generated according to the two tables.
Hint: $P(\hat{Y}(X_1) \neq Y) = P(\hat{Y}(X_1) \neq Y, X_1 = 0) + P(\hat{Y}(X_1) \neq Y, X_1 = 1)$.

★ **SOLUTION:**

$$\begin{aligned} P(\hat{Y}(X_1) \neq Y) &= P(\hat{Y}(X_1) \neq Y, X_1 = 0) + P(\hat{Y}(X_1) \neq Y, X_1 = 1) \\ &= P(Y=1, X_1=0) + P(Y=1, X_1=1) \\ &= 0.06 + 0.14 \\ &= 0.2 \end{aligned}$$

5.2 Naive Bayes with Conditional Independence Violation

Consider two variables $X_1, X_2 \in \{0, 1\}$ and label $Y \in \{0, 1\}$. Y and X_1 are still generated according to Table 1 and Table 2, and then X_2 is created as a **duplicated copy** of X_1.

Question [6 pts]: Now suppose we have trained a Naive Bayes classifier, using *infinite* training data that are generated according to Table 1, Table 2 and the duplication rule. In Table 4, please write down the predictions from the trained Naive Bayes for different configurations of (X_1, X_2). For probability terms in the table, you can write down just the calculation process (e.g., one entry might be $0.8 * 0.3 * 0.3 = 0.072$, and you can just write down $0.8 * 0.3 * 0.3$ to save some time). Hint: the Naive Bayes classifier **does** assume that X_2 is conditionally independent of X_1 given Y.

Table 4: Predictions from the trained Naive Bayes

	$\hat{P}(X_1, X_2, Y = 0)$	$\hat{P}(X_1, X_2, Y = 1)$	$\hat{Y}(X_1, X_2)$
$X_1 = 0, X_2 = 0$	$0.8 \times 0.7 \times 0.7$	$0.2 \times 0.3 \times 0.3$	$\hat{Y} = 0$
$X_1 = 1, X_2 = 1$	$0.8 \times 0.3 \times 0.3$	$0.2 \times 0.7 \times 0.7$	$\hat{Y} = 1$
$X_1 = 0, X_2 = 1$	$0.8 \times 0.7 \times 0.3$	$0.2 \times 0.3 \times 0.7$	$\hat{Y} = 0$
$X_1 = 1, X_2 = 0$	$0.8 \times 0.3 \times 0.7$	$0.2 \times 0.7 \times 0.3$	$\hat{Y} = 0$

★ **SOLUTION:** The naive Bayes model learned from infinite data will have $\hat{P}(Y)$ and $\hat{P}(X_1|Y)$ estimated exactly as Table 1 and Table 2. However, it also has $\hat{P}(X_2|Y)$ incorrectly estimated as Table 2. The resulting predictions are shown in Table 4.

Question [3 pts]: What is the expected error rate of this Naive Bayes classifier on testing examples that are generated according to Table 1, Table 2 and the duplication rule?

★ **SOLUTION:** Note that the testing examples are generated according to the true distribution (i.e., where X_2 is a duplication). We have:

$$\begin{aligned} P(\hat{Y}(X_1, X_2) \neq Y) &= P(\hat{Y}(X_1, X_2) \neq Y, X_1 = X_2 = 0) + P(\hat{Y}(X_1, X_2) \neq Y, X_1 = X_2 = 1) \\ &= P(Y = 1, X_1 = X_2 = 0) + P(Y = 0, X_1 = X_2 = 1) \\ &= P(Y = 1, X_1 = 0) + P(Y = 0, X_1 = 1) \\ &= 0.06 + 0.24 \\ &= 0.3 \end{aligned}$$

Question [3 pts]: Compared to the scenario without X_2, how does the expected error rate change (i.e., increase or decrease)? In Table 4, the decision rule $\hat{Y}$ on **which** configuration is responsible to this change? What actually happened to this decision rule? (You need to *briefly* answer: increase or decrease, the responsible configuration, and what happened.)

★ **SOLUTION:** The expected error rate increases from 0.2 to 0.3, due to the incorrect decision $\hat{Y} = 1$ on the configuration $X_1 = X_2 = 1$. Basically the naive Bayes model makes the incorrect conditional independence assumption and considers both $X_1 = 1$ and $X_2 = 1$ as evidence.

5.3 Logistic Regression with Conditional Independence Violation

Question [2 pts]: Will logistic regression suffer from having an additional variable X_2 that is actually a duplicate of X_1? Intuitively, why (hint: model assumptions)?

★ **SOLUTION:** No. Logistic regression does not make conditional independence assumption. (Note: in the class we did derive the form $P(Y|X)$ of logistic regression from naive Bayes assumptions, but that does not mean logistic regression makes the conditional independence assumption).

Now we will go beyond the intuition. We have a training set $\mathbf{D}_1$ of L examples $\mathbf{D}_1 = \{(X_1^1, Y^1), \ldots, (X_1^L, Y^L)\}$. Suppose we generate another training set $\mathbf{D}_2$ of L examples $\mathbf{D}_2 = \{(X_1^1, X_2^1, Y^1), \ldots, (X_1^L, X_2^L, Y^L)\}$, where in each example X_1 and Y are the same as in $\mathbf{D}_1$ and then X_2 is a duplicate of X_1. Now we learn a logistic regression from $\mathbf{D}_1$, which should contain two parameters: w_0 and w_1; we also learn another logistic regression from $\mathbf{D}_2$, which should have three parameters: w_0', w_1' and w_2'.

Question [4 pts] : First, write down the training rule (maximum conditional likelihood estimation) we use to estimate (w_0, w_1) and (w_0', w_1', w_2') from data. Then, given the training rule, what is the relationship between (w_0, w_1) and (w_0', w_1', w_2') we estimated from $\mathbf{D}_1$ and $\mathbf{D}_2$? Use this fact to argue whether or not the logistic regression will suffer from having an additional duplicate variable X_2.

★ **SOLUTION:**

The training rule for (w_0, w_1) is to maximize:

$$\ln \prod_{l=1}^{L} P(Y^l | X_1^l, w_0, w_1) = \sum_{l=1}^{L} Y^l (w_0 + w_1 X_1^l) - \ln(1 + \exp(w_0 + w_1 X_1^l))$$

The training rule for (w_0', w_1', w_2') is to maximize:

$$\ln \prod_{l=1}^{L} P(Y^l | X_1^l, X_2^l, w_0', w_1', w_2') = \sum_{l=1}^{L} Y^l (w_0' + w_1' X_1^l + w_2' X_2^l) - \ln(1 + \exp(w_0' + w_1' X_1^l + w_2' X_2^l))$$

Since X_2 is a duplication of X_1, the training rule for (w_0', w_1', w_2') becomes maximizing:

$$\begin{aligned} & \sum_{l=1}^{L} Y^l (w_0' + w_1' X_1^l + w_2' X_1^l) - \ln(1 + \exp(w_0' + w_1' X_1^l + w_2' X_1^l)) \\ = & \sum_{l=1}^{L} Y^l (w_0' + (w_1' + w_2') X_1^l) - \ln(1 + \exp(w_0' + (w_1' + w_2') X_1^l)) \end{aligned}$$

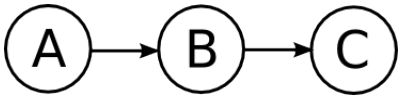
which is basically the same as the training rule for (w_0, w_1), with substitution $w_0 = w'_0$ and $w_1 = w'_1 + w'_2$. This is also the relationship between (w_0, w_1) and (w'_0, w'_1, w'_2) we estimated from $\mathbf{D}_1$ and $\mathbf{D}_2$. As a result, logistic regression will simply split the weight w_1 into $w'_1 + w'_2 = w_1$ when facing duplicated variable $X_2 = X_1$.

6 [Extra Credit 6 pts] Violated assumptions (Grading: Carl Doersch)

Extra Credit Question: This question is optional – do not attempt it until you have completed the rest of the exam. It will not affect the grade curve for the exam, though you will receive extra points if you answer it.

Let A, B, and C be boolean random variables governed by the joint distribution $P(A, B, C)$. Let D be a dataset consisting of n data points, each of which is an independent draw from $P(A, B, C)$, where all three variables are fully observed.

Consider the following Bayes Net, which does not necessarily capture the correct conditional independencies in $P(A, B, C)$.



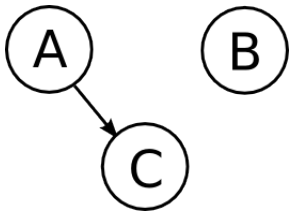
Let $\hat{P}$ be the distribution learned after this Bayes net is trained using D. Show that for any number ϵ, $0 < \epsilon \leq 1$, there exists a joint distribution $P(A, B, C)$ such that $P(C = 1|A = 1) = 1$, but such that the Bayes net shown above, when trained on D, will (with probability 1) learn CPTs where:

$$\hat{P}(C = 1|A = 1) = \sum_{b \in \{0,1\}} \hat{P}(C = 1|B = b)\hat{P}(B = b|A = 1) \leq \epsilon$$

as $|D|$ approaches ∞. Assume that the Bayes net is learning on the basis of the MLE.

You should solve this problem by defining a distribution with the above property. Your final solution may be either in the form of a fully specified joint distribution (i.e. you write out the probabilities for each assignment of the variables A, B, and C), or in the form of a Bayes net with fully specified CPTs. (Hint: the second option is easier.)

★ SOLUTION:



Let $P(A = 1) = \epsilon$, $P(C = 1|A = 1) = 1$, $P(C = 1|A = 0) = 0$. $P(B)$ can be any arbitrary value, as for all values of B, the Bayes net will estimate $P(C = 1|B = b) = \epsilon$.