10-601 Machine Learning, Midterm Exam: Spring 2009 SOLUTION

March 4, 2009

- Please put your name at the top of the table below.
- 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.
- This exam is open book, open notes, no applets, no wireless communication.
- **There are 80 points total on the entire exam, and you have 80 minutes to complete it. Good luck!**

Name:			
Q	Topic	Max. Score	Score
1	Short Answers	15	
2	Decision Trees, Cross Validation, MLE, MAP	16	
3	Drawing Data Sets	10	
4	Bayesian Networks	20	
5	Decision Boundaries for Gaussian Naive Bayes	12	
6	Maximum Likelihood Estimation	7	
Total		80	

1 Short Answers [15 points]

A. (3 points) Give a *one sentence* reason why:

- we might prefer Decision Tree learning over Logistic Regression for a particular learning task.

  $\star$ *Solution*: If we want our learner to produce rules easily interpreted by humans.

- we might prefer Logistic Regression over Naive Bayes for a particular learning task.

  $\star$ *Solution*: If we know that the conditional independence assumptions made by Naive Bayes are not true for our problem, and we have lots of training data.

- we choose parameters that minimize the sum of squared training errors in Linear Regression.

  $\star$ *Solution*: Because this corresponds to the MLE assuming that data is generated from a linear function plus Gaussian noise.

B. (3 points) Suppose we train several classifiers to learn $f : X \rightarrow Y$, where X is the feature vector $X =< X_1, X_2, X_3 >$. Which of the following classifiers contains sufficient information to allow calculating $P(X_1, X_2, X_3, Y)$? If you answer yes, give a brief sketch of how. If you answer no, state what is missing.

- Gaussian Naive Bayes

  $\star$ *Solution*: Yes, we can estimate $P(X_1, X_2, X_3, Y) = P(Y)P(X_1|Y)P(X_2|Y)P(X_3|Y)$.

- Logistic Regression

  $\star$ *Solution*: No, we cannot compute $P(X)$.

- Linear Regression

  $\star$ *Solution*: No, we cannot compute $P(X)$.

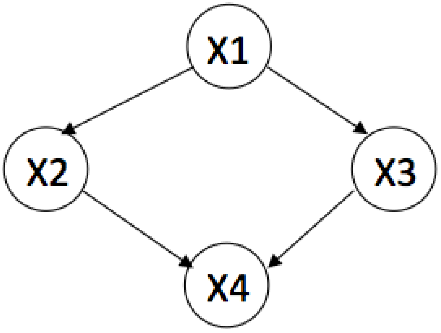
C. (3 points) In class we defined *conditional independence* by saying that random variable X is conditionally independent of Y given Z if and only if:

$$P(X|Y, Z) = P(X|Z) \tag{1}$$

Prove that if $P(XY|Z) = P(X|Z)P(Y|Z)$, then X is conditionally independent of Y given Z (*hint: this is a two-line proof*).

$\star$ *Solution*: Assume $P(XY|Z) = P(X|Z)P(Y|Z)$. $P(XY|Z) = P(X|Y,Z)P(Y|Z)$ (by the Chain Rule), so $P(X|Z)P(Y|Z) = P(X|Y,Z)P(Y|Z)$ (substituting our assumption). Dividing both sides by $P(Y|Z)$ yields that $P(X|Z) = P(X|Y,Z)$, so we have our proof from the stated definition of conditional independence.

D. (6 points) Consider the Bayes network below, defined over four Boolean variables.



- How many parameters are needed to define $P(X1, X2, X3, X4)$ for this Bayes Net?

  ⋆ *Solution*: 9

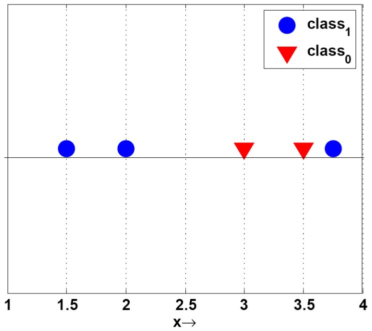
- Give the formula that calculates $P(X1 = 1, X2 = 0, X3 = 1, X4 = 0)$ using only the Bayes net parameters. Use notation like $P(X1 = 0|X2 = 1, X4 = 0)$ to refer to each Bayes net parameter you use in your formula.

  ⋆ *Solution*: $P(X1 = 1)P(X2 = 0|X1 = 1)P(X3 = 1|X1 = 1)P(X4 = 0|X2 = 0, X3 = 1)$

- Give the formula that calculates $P(X1 = 1, X4 = 0)$ using only the Bayes net parameters.

  ⋆ *Solution*: $P(X1 = 1)P(X2 = 1|X1 = 1)P(X3 = 1|X1 = 1)P(X4 = 0|X2 = 1, X3 = 1) + P(X1 = 1)P(X2 = 1|X1 = 1)P(X3 = 0|X1 = 1)P(X4 = 0|X2 = 1, X3 = 0) + P(X1 = 1)P(X2 = 0|X1 = 1)P(X3 = 1|X1 = 1)P(X4 = 0|X2 = 0, X3 = 1) + P(X1 = 1)P(X2 = 0|X1 = 1)P(X3 = 0|X1 = 1)P(X4 = 0|X2 = 0, X3 = 0)$

2 Decision Trees, Cross Validation, MLE and MAP [16 points]



Consider the 1-dimensional data set shown above, based on the single real-valued attribute x. Notice there are two classes (values of Y), and five data points.

We will learn Decision Trees from this data using the ID3 algorithm. Given real-valued attributes, ID3 considers splitting the possible values of the attribute into two sets based on a threshold (i.e., given a real valued attribute such as x, ID3 considers tests of the form $x > t$ where t is some threshold value). It considers alternative thresholds (data splits), and it selects among these using the same information gain heuristic that it uses for other discrete-valued attributes.

Given n training examples, ID3 considers exactly $n + 1$ possible values for the threshold. In particular, for each pair of adjacent adjacent training examples of x, it considers the threshold value midway between the two examples. In addition, it considers the threshold just to the right of the largest training example, and just to the left of the lowest training value.

Important: Assume that if your Decision Tree learner finds that two data splits $x < t_1$ and $x < t_2$ have the same information gain, then it breaks ties by picking the leftmost threshold. (e.g., if $t_1 < t_2$ then it picks the split $(x < t_1)$).

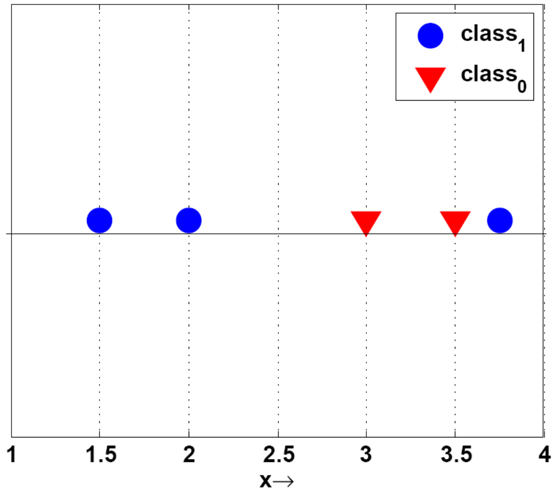
Let algorithm $DT1$ be the algorithm that learn a decision tree with only one boolean split (a depth 1 tree). Let DT^* be the algorithm that learns a decision tree with as many boolean splits as necessary to perfectly classify the training data (using a different threshold for each tree node).

A. (4 points) What is the training set error of $DT1$ on the data? What is its leave one out cross validation error?

$\star$ *Solution*: The training set error will be 1. It's LOOCV error will be 1.

B. (4 points) What is the training set error of DT^* on the data? What is its leave one out cross validation error?

$\star$ *Solution*: The training set error will be 0. It's LOOCV error will be 2.



Now consider a new class of decision trees where leaves have probabilistic labels. Each leaf node gives the probability of each possible label, where the probability is the fraction of points at that leaf node with that label. For example, a decision tree learned from the data set above with zero splits would say $P(Y = 1) = 3/5$ and $P(Y = 0) = 2/5$. A decision tree with one split (at $x = 2.5$) would say $P(Y = 1) = 1$ if $x < 2.5$, and $P(Y = 1) = 1/3$ if $x \geq 2.5$.

C. (3 points) For the above data set, draw a tree that maximizes the likelihood of the data.

⋆ *Solution*: A correct tree will label the data perfectly. One correct answer would be to split at 2.5, then at 3.625, with the appropriate labels.

D. (5 points) Consider a prior probability distribution $P(T)$ over trees that penalizes the number of splits in the tree.

$$P(T) \propto \left(\frac{1}{4}\right)^{splits(T)^2}$$

where T is a tree, $splits(T)$ is the number of splits in T, and $\propto$ means "is proportional to".

For the same data set, give the MAP tree when using this prior, $P(T)$, over trees. Show your work. The posterior probabilities of each possible tree should turn out to be fractions– you should not need a calculator to select the maximum posterior.

⋆ *Solution*: A tree with zero splits will say $P(Y = 1) = 3/5$, and the prior $P(T)$ on such a tree will be proportional to 1. So the posterior probability of such a tree will be proportional to $(3/5)^3(2/5)^2 = 36/3125$.

The best tree with one split will split at $x = 2.5$. The posterior probability of the data will be proportional to $(2/3)^2(1/3)(1/4) = 1/27$.

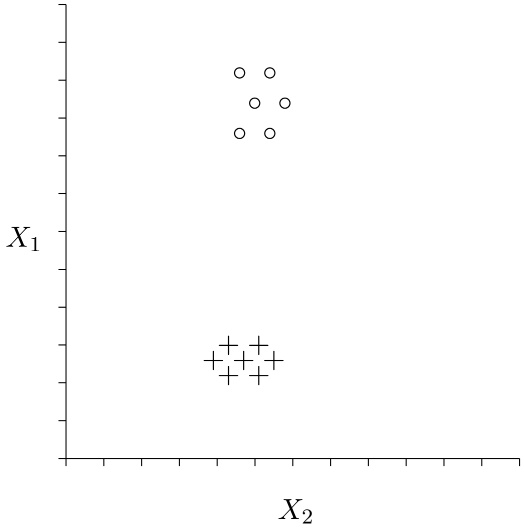
The tree with two splits will perfectly label the data, but $P(T)$ will be proportional to $(1/4)^4 = 1/256$, so the posterior will be proportional to $1/256$.

Thus the tree with one split has the highest posterior and is the MAP tree.

3 Drawing Data Sets [10 points]

In this problem, you will draw data sets with certain properties. These data sets have two real-valued features, X_1 and X_2, and two classes, $Y = 0$ and $Y = 1$. When drawing a data point, please indicate the label by drawing a small circle ($\circ$) for the label $Y = 0$ and a small plus sign (+) for the label $Y = 1$.

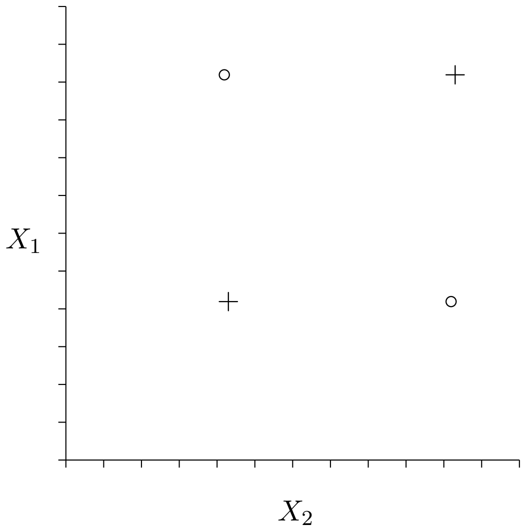
As an example, here is a data set that can be perfectly classified by a decision tree with a single split, as well as by logistic regression:



Consider the class of decision trees of depth two. Let us call this class $DT2$. It can have two levels of splits, and therefore four leaf nodes. Consider also the class of logistic regression models, which we will call LR.

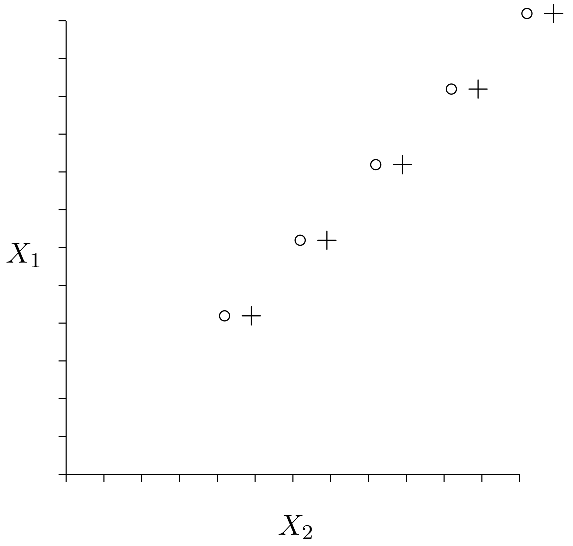
A. (4 points) Draw a data set where a hypothesis exists in $DT2$ that can perfectly classify the data, but no such hypothesis exists in LR.

$\star$ *Solution*: Any answer which could be learned by a depth two decision tree but wasn't linearly separable is OK. XOR is such an example. A tree could split X1 once and then X2 one more on each side to properly label this data set, but no line exists that can separate the two classes:


X_1
X_2

B. (3 points) Draw a data set where a hypothesis exists in LR that can perfectly classify the data, but no such hypothesis exists in $DT2$.

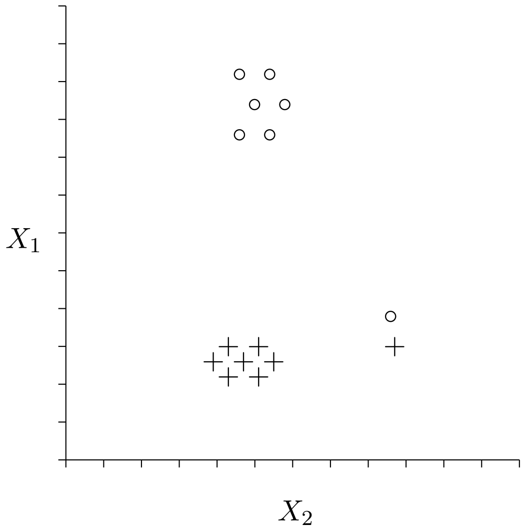
⋆ *Solution*: Since the decision tree must make splits that are aligned with the axes, a data set with several points that could be separated by a diagonal line works here:



C. (3 points) Now consider a Gaussian Naive Bayes model where the variance is assumed to be independent of both X_i and Y. That is, there is only one (shared) variance parameter, σ.

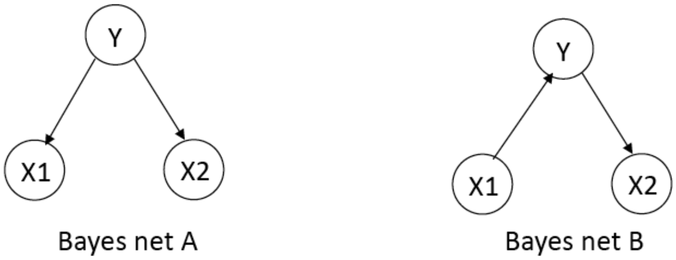
Draw a data set of training data where unregularized Logistic Regression will *learn a hypothesis* which perfectly classifies the training data, while this Gaussian Naive Bayes model will not.

⋆ *Solution*: GNB will learn the means of the data, and regardless of the learned value of σ, the decision boundary will be a line that runs exactly between the midpoint of the two means, and is perpendicular to the line that connects the means. A sufficiently skewed data set that is linearly separable makes it so that LR will learn to separate the data sets, while the means learned by GNB will result in a line that does not:



4 Bayesian Networks [20 points]

Consider the two Bayesian networks below defined over three Boolean random variables. Notice the only difference in their graphs lies in the arrow between Y and X_1.



A. (2 points) Of course these two Bayes nets both describe a joint probability distribution $P(X1, X2, Y)$. However, they factor this joint distribution into different forms. Please give the form for each:

```
Bayes net A:  P(X1,X2,Y) =
```

$\star$ *Solution*: $P(Y)P(X_1|Y)P(X_2|Y)$

```
Bayes net B: P(X1,X2,Y) =
```

$\star$ *Solution*: $P(X_1)P(Y|X_1)P(X_2|Y)$

B. (4 points) Now consider the joint probability distribution $P(X1, X2, Y)$ that assigns probability 1 to the joint assignment $X1 = 1, X2 = 1, Y = 1$, and assigns probability zero to the other seven possible assignments. Write down the CPT's for Bayes network A that would define this joint distribution (just this one network).

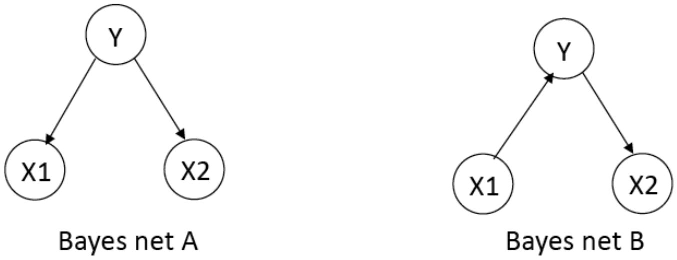
$\star$ *Solution*: $P(Y = 1) = 1, P(X_1 = 1|Y = 1) = 1, P(X_2 = 1|Y = 1) = 1$. The other two CPT entries do not matter since $P(Y = 0) = 0$.

C. (4 points) Describe a joint probability distribution that *can NOT* be represented by Bayes network A. Most importantly, give a one-sentence explanation of why it cannot.

$\star$ *Solution*:

```
P(X1=0, X2=0, Y=0)=0
P(X1=0, X2=0, Y=1)=0
P(X1=0, X2=1, Y=0)=0
P(X1=0, X2=1, Y=1)=1
P(X1=1, X2=0, Y=0)=0
P(X1=1, X2=0, Y=1)=1
P(X1=1, X2=1, Y=0)=0
P(X1=1, X2=1, Y=1)=0
```

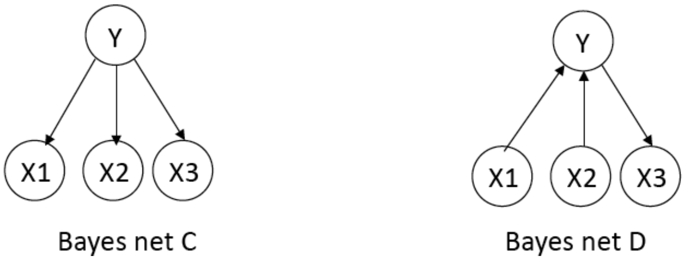
X_1 and X_2 are not conditionally independent given Y, so Bayes network A cannot represent this distribution.



D. (5 points) Do Bayes Net A and Bayes Net B above represent different conditional independence relations? If so, please give a condition independence assumption that holds for one of these networks but not the other. If they represent the same conditional independence assumptions, then list the complete set of conditional independencies they represent.

⋆ *Solution*: They represent the same conditional independence relations: X_1 is C.I. of X_2 given Y.

E. (5 points) True or false: The set of joint probability distributions $P(X1, X2, X3, Y)$ that can be represented by Bayes Network C below is identical to the set of joint distributions that can be represented by Bayes Network D below. [One point for the correct answer. Four points for a lucid, two sentence explanation].



⋆ *Solution*: False. The networks have different CI assumptions. In Bayes Net C, X_1 is C.I. of X_2 given Y, while in Bayes Net D, this is not true.

5 The Decision Boundary of Gaussian Naive Bayes [12 points]

Consider Gaussian Naive Bayes classifiers for a dataset with a single attribute x and two classes 0 and 1. We will classify a point $< x >$ as class 1 if

$$\begin{aligned} P(y=1|x) &\geq P(y=0|x) \\ ln\frac{P(y=1|x)}{P(y=0|x)} &\geq 0 \end{aligned} \tag{2}$$

In class you have seen that for a *certain family* of Gaussian Naive Bayes (GNB) models, the above decision rule gives rise to a linear decision boundary. In the following question you will be given different GNB models, and you will write down the expression for the corresponding decision boundaries. We will use $N(\mu, \sigma^2)$ to denote a normal distribution with mean μ and variance σ^2. *Hint: You should be able to find the answer by drawing the distributions. You will not need to derive equations.*

Consider the following Gaussian Naive Bayes classifier (model 3).

$$\begin{aligned} x|y=0 &\sim N(0,1) \\ x|y=1 &\sim N(2,1) \\ P(y=1) &= 0.5 \end{aligned} \tag{3}$$

A. (3 points) Is the decision boundary of this GNB classifier (model 3) linear? In other words, can you write a linear expression of the form $w_0 + w_1 x \geq 0$ to represent the decision boundary of this GNB model? If yes, please give the values of w_0 and w_1. If no, please explain why not.

$\star$ *Solution*: Yes, the decision boundary is linear. It classifies a point (x) as 1 if $x \geq 1$, and 0 otherwise. Hence the coefficients are $w_0 = -k$, and $w_1 = k$, for any $k > 0$.

Now consider a different GNB classifier (model 5). The new parameters for the two Gaussian distributions are

$$\begin{aligned} x|y=0 &\sim N(0,1/4) \\ x|y=1 &\sim N(0,1) \\ P(y=1) &= 0.5 \end{aligned} \tag{4}$$

B. (4 points) Is the decision boundary of this GNB classifier (model 4) linear? If yes, please check the right decision boundary from the following options. If you check option (e), give a brief explanation.

(a) Classify as class 1 if $x \geq 1/2$.

(b) Classify as class 1 if $x \leq -1/2$.

(c) Classify as class 1 if $x \leq 1$.

(d) Classify as class 1 if $x \geq -1$.

(e) Decision boundary is not linear.

⋆ *Solution*: No. The decision boundary is not linear. Hence option (e) is correct. The class distributions have different variances. Hence we will have a quadratic decision boundary.

C. (5 points) Now consider a quadratic decision boundary. In a quadratic decision boundary we add a new feature x^2 for a data-point $< x, y >$. Thus we convert every data-point $< x, y >$ into $< x^2, x, y >$. A linear decision boundary for this modified dataset yields a quadratic boundary for the original dataset. Is it possible to find a quadratic decision boundary which matches the decision boundary of the GNB model from (4) exactly? If yes, please check the right decision boundary from the following options. If you check option (d), give a brief explanation.

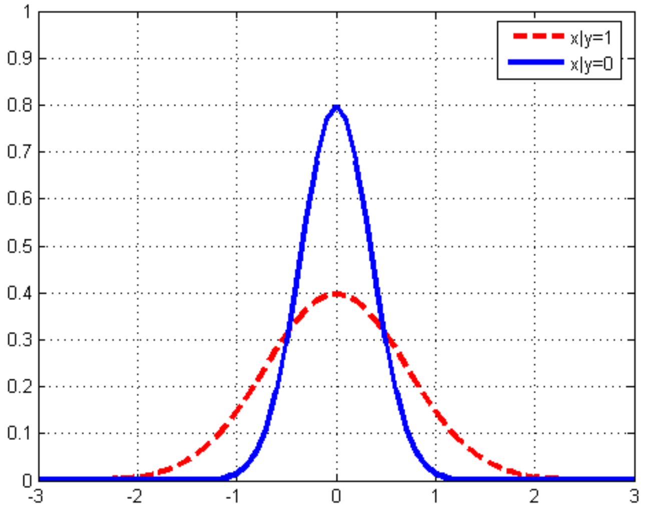
(a) Classify as class 1 if $x \geq 0.48$ or $x \leq -0.48$

(b) Classify as class 1 if $x \leq 0.95$ and $x \geq -0.95$

(c) Classify as class 1 if $x \leq 0.48$ and $x \geq -0.48$

(d) Decision boundary is not quadratic.

⋆ *Solution*: Option (a) is correct. A simple way of seeing this is to draw the gaussian distributions (see figure below). We see that the gaussian for $< x|y = 1 >$ (dashed line) dominates that for $< x|y = 0 >$ in the region roughly matching with option (a).



6 Maximum Likelihood Estimation [7 points]

You are given a dataset with N records in which the i^{th} record has a real-valued input attribute x_i and a real-valued output y_i, which is generated from a Gaussian distribution with mean $\sin(wx_i)$, and variance 1.

$$P(y_i|w, x_i) = \frac{1}{\sqrt{2\pi}} \exp \frac{-(y_i - \sin(wx_i))^2}{2} \tag{5}$$

We have one unknown parameter w in the above model and we want to learn the maximum likelihood estimate of it from the data.

A. (2 point) Write down the expression for the data likelihood as a function of w and the observed x and y values.

$\star$ *Solution*: The conditional data likelihood is given by $\frac{1}{(2\pi)^{N/2}} \exp[\sum_{i=1}^{N} \frac{-(y_i - \sin(wx_i))^2}{2}]$.

B. (5 points) If you compute the maximum likelihood estimate of the parameter w, which of the following equations is satisfied? If none of them are satisfied, check option (f). *hint:* $d/dw(\sin(wx)) = x\cos(wx)$.

(a) $\sum_i \cos^2(wx_i) = \sum_i y_i \sin(x_i)$

(b) $\sum_i \cos^2(wx_i) = \sum_i y_i \sin(2wx_i)$

(c) $\sum_i \sin(wx_i)\cos(wx_i) = \sum_i y_i \sin(wx_i/2)$

(d) $\sum_i x_i \sin(wx_i)\cos(wx_i) = \sum_i x_i y_i \cos(wx_i)$

(e) $w\sum_i \cos(x_i) = \sum_i x_i y_i \cos(x_i)$

(f) None of the above.

$\star$ *Solution*: Correct option is (d).