# Q1 Probability and MLE [20 pts]

1. (a) Suppose we wish to calculate $P(H|E_1, E_2)$ and we have no conditional independence information. Which of the following sets of numbers are sufficient for the calculation ?

   i. $P(E_1, E_2), P(H), P(E_1|H), P(E_2|H)$
   ii. $P(E_1, E_2), P(H), P(E_1, E_2|H)$
   iii. $P(H), P(E_1|H), P(E_2|H)$

   Bayes' Rule: $P(H|E_1,E_2) = \frac{P(E_1,E_2|H)P(H)}{P(E_1,E_2)}$

   (b) Suppose we know that $P(E_1|H, E_2) = P(E_1|H)$ for all values of $H, E_1, E_2$. Now which of the above three sets are sufficient ?

   (i) because $P(E_1,E_2|H) = P(E_1|H)P(E_2|H)$
   (ii) it just ignores the given independence relations.

2. Which of the following statements are true ? If none of them are true, write NONE.

   (a) If $X$ and $Y$ are independent then $E[2XY] = 2E[X]E[Y]$ and $Var[X + 2Y] = Var[X] + Var[Y]$.
   ~~$Var[2XY] = Var[X] + Var[Y]$~~

   (b) If $X$ and $Y$ are independent and $X > 1$ then $Var[X+2Y^2] = Var[X]+4Var[Y^2]$ and $E[X^2-X] \geq Var[X]$.

   (c) If $X$ are $Y$ are not independent then $Var[X + Y] = Var[X] + Var[Y]$.

   (d) If $X$ and $Y$ are independent then $E[XY^2] = E[X]E[Y]^2$ and $Var[X + Y] = Var[X] + Var[Y]$.

   (e) If $X$ and $Y$ are not independent and $f(X) = X^2$ then $E[f(X)Y] = E[f(X)]E[Y]$ and $Var[X + 2Y] = Var[X] + 4Var[Y]$

   (b)

OVER FOR REASONS →

3. You are playing a game with two coins. Coin 1 has a $\theta$ probability of heads. Coin 2 has a $2\theta$ probability of heads. You flip these coins several times and record your results:

| Coin | Result |
|---|---|
| 1 | Head |
| 2 | Tail |
| 2 | Tail |
| 2 | Tail |
| 2 | Head |

   (a) What is the log-likelihood of the data given $\theta$ ?

$$L(\theta) = P(\text{data}|\theta) = P(\text{coin 1} = \text{Head})\,[P(\text{coin 2} = \text{Tail})]^3\,P(\text{coin 2} = \text{Head})$$
$$= \theta(1-2\theta)^3 2\theta = 2\theta^2(1-2\theta)^3$$
$$\ell(\theta) = \log L(\theta) = \log 2 + 2\log\theta + 3\log(1-2\theta)$$

   (b) What is the maximum likelihood estimate for $\theta$ ?

$$0 = \frac{\partial \ell(\theta)}{\partial \theta} = \frac{2}{\theta} + \frac{3(-2)}{(1-2\theta)} \Rightarrow \frac{2(1-2\theta) - 6\theta}{\theta(1-2\theta)} = 0 \Rightarrow \hat{\theta}_{MLE} = 1/5$$



reminder: $\hat{\theta}_{MLE} = \arg\max_\theta L(\theta) = \arg\max_\theta \ell(\theta)$ b/c $\log(\cdot)$ is monotone! [maximizing $L(\theta)$ directly is hard]

2. Relevant properties

$E[aX] = aE[X] \quad a \in \mathbb{R}$

$Var[aX] = a^2 Var[X] \quad a \in \mathbb{R}$

if $f(X)$ is nonlinear then $E[f(X)] \neq f(E[X])$

If X and Y are independent

$E[XY] = E[X]E[Y]$

$Var[X+Y] = Var[X] + Var[Y]$

$E[X+Y] = E[X] + E[Y]$

If X and Y are not independent

$E[XY] \neq E[X]E[Y]$

$Var[X+Y] \neq Var[X] + Var[Y]$ (cf. $X \sim N(0, \sigma^2)$, $Y = -X$)

$E[X+Y] = E[X] + E[Y]$

These properties are enough to show that (a), (c), (d), (e) are false.

For (b) $E[X^2 - X] = E[X^2] - E[X]$

$Var[X] \triangleq E[(X - E[X])^2]$

$= E[X^2 - 2E[X]X + E[X]^2]$

$= E[X^2] - E[2E[X]X] + E[E[X]^2]$

$= E[X^2] - 2E[X]E[X] + E[X]^2$

$= E[X^2] - E[X]^2$

But since $X \geq 1$ $E[X]^2 \geq E[X]$ and so

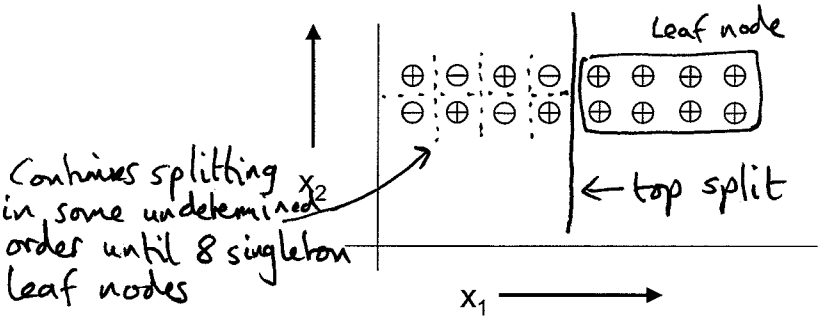
$$E[X^2] - E[X] \geq E[X^2] - E[X]^2$$

$$\| \qquad\qquad \| $$

$$E[X^2 - X] \geq Var[X]$$

## Q2 Decision Trees [20 pts]

1. The figure below shows a dataset with two inputs $X_1$ and $X_2$ and one output $Y$, which can take on the values positive (+) or negative (-). There are 16 datapoints: 12 are positive and 4 are negative.




Answer to (a)

Assume we are testing two extreme decision tree learning algorithms. Algorithm OVERFIT builds a decision tree in the standard fashion, but never prunes. Algorithm UNDERFIT refuses to risk splitting at all, and so the entire decision tree is just one leaf node.

(a) Exactly how many leaf-nodes will be in the decision tree learned by OVERFIT on this data?

9 (see picture above)

(b) What is the leave-one-out classification error of using OVERFIT on our dataset? Report the total number of misclassifications.

Every point in the left half will be misclassified because it will be in a singleton leaf node owned by the opposing class. Every point on right will be fine. Answer = 8

(c) What is the leave-one-out classification error of using UNDERFIT on our dataset? Report the total number of misclassifications.

In all 16 folds, the ⊕ will be the majority class, so only errors will be on ⊖. There are 4 ⊖ nodes. Ans = 4

(d) Now, suppose we are learning a decision tree from a dataset with $M$ binary-valued inputs and $R$ training points. What is the maximum possible number of leaves in the decision tree. Circle one of the following answers:

If $R < 2^M$ then largest tree has a single point at each leaf, i.e. R leaves.

If $R \geq 2^M$ then splitting must stop after all M attributes have been tested. That makes $2^M$ leaves.

$R, \log_2(R), R^2, 2^R, M, \log_2(M), M^2, 2^M,$

$\min(R, M), \min(R, \log_2(M)), \min(R, M^2), \min(R, 2^M),$

$\min(\log_2(R), M), \min(\log_2(R), \log_2(M)), \min(\log_2(R), M^2), \min(\log_2(R), 2^M),$

$\min(R^2, M), \min(R^2, \log_2(M)), \min(R^2, M^2), \min(R^2, 2^M),$

$\min(2^R, M), \min(2^R, \log_2(M)), \min(2^R, M^2), \min(2^R, 2^M),$

$\max(R, M), \max(R, \log_2(M)), \max(R, M^2), \max(R, 2^M),$

$\max(\log_2(R), M), \max(\log_2(R), \log_2(M)), \max(\log_2(R), M^2), \max(\log_2(R), 2^M),$

$\max(R^2, M), \max(R^2, \log_2(M)), \max(R^2, M^2), \max(R^2, 2^M),$

$\max(2^R, M), \max(2^R, \log_2(M)), \max(2^R, M^2), \max(2^R, 2^M)$

Thus answer = $\min(R, 2^M)$

Q3

## Linear Regression

Consider fitting the linear regression model for these data

| x | -1 | 0 | 2 |
|---|---|---|---|
| y | 1 | -1 | 1 |

(b) Fit $Y_i = \beta_0 + \epsilon_i$ (degenerated linear regression), find $\beta_0$.

$\beta_0$ =argmin $\sum(Y_i - \beta_0)^2$

$\beta_0 = 1/3$

(b) Fit $Y_i = \beta_1 X_i + \epsilon_i$ (linear regression without the constant term), find $\beta_0$ and $\beta_1$.

$\beta_1$ =argmin $\sum(Y_i - \beta_1 X_i)^2$

$\beta_1 = \sum X_i Y_i / \sum X_i^2 = 1/5$

## Q4 Conditional Independence [5 pts]

1. Consider the following joint distribution over the random variables A, B, and C.

| A | B | C | P(A,B,C) |
|---|---|---|---|
| 0 | 0 | 0 | 1/8 |
| 0 | 1 | 0 | 1/8 |
| 0 | 0 | 1 | 1/8 |
| 0 | 1 | 1 | 1/8 |
| 1 | 0 | 0 | 1/8 |
| 1 | 1 | 0 | 1/8 |
| 1 | 0 | 1 | 1/8 |
| 1 | 1 | 1 | 1/8 |

(a) True or False: A is conditionally independent of B given C.

True, because $\forall_{ijk}\ P(A=i|B=j,C=k) = P(A=i|C=k)$

(b) If you answered part (a) with TRUE, make a change to the top two rows of this table to create a joint distribution in which the answer to (a) is FALSE.

If you answered part (a) with FALSE, make a change to the top two rows of this table to create a joint distribution in which the answer to (a) is TRUE.

One possible change is

| A | B | C | P(ABC) |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 1/4 |

note any change made to these two rows must still result in the table representing a joint probability distribution whose probabilities sum to one.

## Q5 Generative vs Discriminative Classifiers [15 pts]

1. You wish to train a classifier to predict the gender (a boolean variable, $G$) of a person based on that person's weight (a continuous variable, $W$) and whether or not they are a graduate student (a boolean variable, $S$). Assume that $W$ and $S$ are conditionally independent given $G$. Also, assume that the variance of the probability distribution $P(Weight|Gender = female)$ equals the variance for $P(Weight|Gender = male)$.

   (a) Is it reasonable to train a Naive Bayes classifier for this task?

   Yes. W and S are conditionally independent given G.

   (b) If not, explain why not, and describe how you might reformulate this problem to allow training a naive Bayes classifier. If so, list every probability distribution your classifier must learn, what form of distribution you would use for each, and give the total number of parameters your classifier must estimate from the training data.

   We must estimate 6 parameters:

   P(G) Bernoulli → $P(G=1) \equiv \pi$ (note $P(G=0)$ need not be estimated separately. It is $1 - P(G=1)$)

   P(S|G) Bernoulli → $P(S=1 \mid G=1) \equiv \theta_1$

   Bernoulli → $P(S=1 \mid G=0) \equiv \theta_0$

   P(W|G) Normal →
   - $\sigma_W$ - variance for the Normal distributions governing W
   - $\mu_{W|G=1}$ - mean for $P(W|G=1)$
   - $\mu_{W|G=0}$ - mean for $P(W|G=0)$

   (c) Note one difference between the above $P(Gender|Weight, Student)$ problem and the problems we discussed in class is that the above problem involves training a classifier over a *combination* of boolean and continuous inputs. Now suppose you would like to train a discriminative classifier for this problem, to directly fit the parameters of $P(G|W, S)$, under the conditional independence assumption. Assuming that W and S are conditionally independent given G, is it correct to assume that $P(G = 1|W, S)$ can be expressed as a conventional logistic function:

$$P(G = 1|W, S) = \frac{1}{1 + \exp(w_0 + w_1 W + w_2 S)}$$

   If not, explain why not. If so, prove this.

   Yes. This can be shown by combining the derivation in Tom's Naive Bayes chapter draft (which covers the case of Normal variables) with the solution to a question from homework 2 (which covers Boolean variables).

from eq 19 in Tom's handout, using our variables G, W, + S, we have:

$$P(G=1|WS) = \frac{1}{1 + \exp\left(\ln\frac{1-\pi}{\pi} + \ln\frac{P(W|G=0)}{P(W|G=1)} + \ln\frac{P(S|G=0)}{P(S|G=1)}\right)}$$

from Tom's handout this equals: $W\left(\frac{\mu_0 - \mu_1}{\sigma^2}\right) + \left(\frac{\mu_1^2 - \mu_0^2}{2\sigma^2}\right)$

from HW2, this is: $S \ln\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)} + \ln\frac{1-\theta_0}{1-\theta_1}$

therefore:

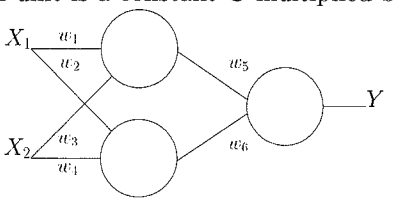
$$w_0 = \ln\frac{1-\pi}{\pi} + \ln\frac{1-\theta_0}{1-\theta_1} + \frac{\mu_1^2 - \mu_0^2}{2\sigma^2}$$

$$w_1 = \frac{\mu_0 - \mu_1}{\sigma^2}$$

$$w_2 = \ln\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}$$

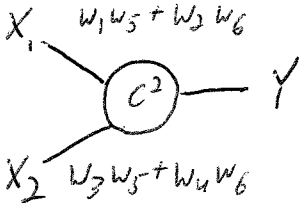
# Q6 Neural Networks [20 pts]

1. For this question, suppose we have a Neural Network (shown below) with linear activation units. In other words, the output of each unit is a constant **C** multiplied by the weighted sum of inputs.




(a) Can any function that is represented by the above network also be represented by a single unit ANN (or perceptron). If so, draw the equivalent perceptron, detailing the weights and the activation function. Otherwise, explain why not.

Yes




This answer uses $c^2$ as the activation function.
Any answer that provided equivalent function by a correct linear combination of weights was acceptable.

(b) Can the space of functions that is represented by the above ANN also be represented by linear regression? (Yes/No)

Yes Any function in the network above has the form:

$$Y = \underbrace{c^2(w_1w_5 + w_2w_6)}_{\beta_1} X_1 + \underbrace{c^2(w_3w_5 + w_4w_6)}_{\beta_2} X_2$$

This is a linear regression on $X_1, X_2$ with coefficients $\beta_1, \beta_2$.

2. Consider the XOR function: $Y = (X_1 \wedge \neg X_2) \vee (\neg X_1 \wedge X_2)$. We can also express this as:

$$Y = \begin{cases} > \frac{1}{2} & X_1 \neq X_2 \\ < \frac{1}{2} & otherwise \end{cases}$$

It is well known that XOR cannot be implemented by a single perceptron. Draw a fully connected three unit ANN that has binary inputs $X_1$, $X_2$, 1 and output $Y$.

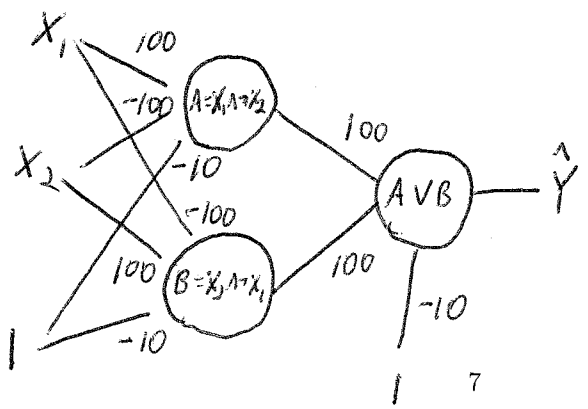
Select weights that implement $\mathbf{Y = (X_1 \text{ XOR } X_2)}$.

For this question, assume the sigmoid activation function:

$$y = \frac{1}{1 + \exp(-(w_0 + w_1x_1 + w_2x_2))}$$

We need to implement the following truth table:

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |




We use the decomposition above

$$Y = \underbrace{(X_1 \wedge \neg X_2)}_{A} \vee \underbrace{(\neg X_1 \wedge X_2)}_{B}$$

The first layer implements A & B
The last node implements "OR"

- keep in mind: weighted sum is (-) negative => sigmoid < .5 Otherwise sigmoid > .5
- If relative magnitude of weights is skewed, the output may also be skewed.