10-601 Machine Learning: Homework 2

Due 5 p.m. Wednesday, February 4, 2015

Problem 1: More Probability Review

(a) [**4 Points**] For events A and B, prove

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

Solution: This is Bayes theorem. From the chain rule, we have that $P(A,B) = P(A|B)P(B)$ and $P(A,B) = P(B|A)P(A)$. Therefore, $P(A|B)P(B) = P(B|A)P(A)$. Rearranging gives the desired equality.

(b) [**4 Points**] For events A, B, and C, rewrite $P(A,B,C)$ as a *product* of several conditional probabilities and one unconditional probability involving a single event. Your conditional probabilities can use only one event on the left side of the conditioning bar. For example, $P(A|C)$ and $P(A)$ would be okay, but $P(A,B|C)$ is not.

Solution: $P(A,B,C) = P(A|B,C)P(B|C)P(C)$. This follows from repeated application of the chain rule:

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

(c) [**4 Points**] Let A be any event, and let X be a random variable defined by

$$X = \begin{cases} 1 & \text{if event } A \text{ occurs} \\ 0 & \text{otherwise.} \end{cases}$$

X is sometimes called the indicator random variable for the event A. Show that $\mathbb{E}[X] = P(A)$, where $\mathbb{E}[X]$ denotes the *expected value* of X.

Solution:

$$\begin{aligned} \mathbb{E}[X] &= 0 \cdot P(X=0) + 1 \cdot P(X=1) \\ &= 0 \cdot P(\text{not } A) + 1 \cdot P(A) \\ &= P(A). \end{aligned}$$

(d) Let X, Y, and Z be random variables taking values in $\{0,1\}$. The following table lists the probability of each possible assignment of 0 and 1 to the variables X, Y, and Z:

	$Z=0$		$Z=1$	
	$X=0$	$X=1$	$X=0$	$X=1$
$Y=0$	$1/15$	$1/15$	$4/15$	$2/15$
$Y=1$	$1/10$	$1/10$	$8/45$	$4/45$

For example, $P(X=0, Y=1, Z=0) = 1/10$ and $P(X=1, Y=1, Z=1) = 4/45$.

(i) [**4 Points**] Is X independent of Y? Why or why not?

(ii) [**4 Points**] Is X conditionally independent of Y given Z? Why or why not?

(iii) [**4 Points**] Calculate $P(X = 0|X + Y > 0)$.

Solution: Using the facts that for any events A and B, (i.e., any subsets of the possible assignments of 0 and 1 to the variables X, Y, and Z) we have $P(A) = \sum_{(x,y,z)\in A} P(X = x, Y = y, Z = z)$ and $P(A|B) = P(A,B)/P(B)$, we have the following solutions

(i) No.

$$P(X = 0) = 1/15 + 1/10 + 4/15 + 8/45 = 11/18,$$
$$P(Y = 0) = 1/15 + 1/15 + 4/15 + 2/15 = 8/15,$$

and

$$P(X = 0|Y = 0) = \frac{P(X = 0, Y = 0)}{P(Y = 0)} = \frac{1/15 + 4/15}{8/15} = 5/8.$$

Since $P(X = 0)$ does not equal $P(X = 0|Y = 0)$, X is not independent of Y.

(ii) For all pairs $y, z \in \{0,1\}$, we need to check that $P(X = 0|Y = y, Z = z) = P(X = 0|Z = z)$. That the other probabilities are equal follows from the law of total probability. First we have

$$P(X = 0|Y = 0, Z = 0) = \frac{1/15}{1/15 + 1/15} = 1/2$$
$$P(X = 0|Y = 1, Z = 0) = \frac{1/10}{1/10 + 1/10} = 1/2$$
$$P(X = 0|Y = 0, Z = 1) = \frac{4/15}{4/15 + 2/15} = 2/3$$
$$P(X = 0|Y = 1, Z = 1) = \frac{8/45}{8/45 + 4/45} = 2/3.$$

Second

$$P(X = 0|Z = 0) = \frac{1/15 + 1/10}{1/15 + 1/15 + 1/10 + 1/10} = 1/2$$
$$P(X = 0|Z = 1) = \frac{4/15 + 8/45}{4/15 + 2/15 + 8/45 + 4/45} = 2/3.$$

This shows that X is independent of Y given Z.

(iii)

$$P(X = 0|X + Y > 0) = \frac{1/10 + 8/45}{1/15 + 1/10 + 1/10 + 2/15 + 4/45 + 8/45} = 5/12.$$

Problem 2: Maximum Likelihood and Maximum a Posteriori Estimation

This problem explores two different techniques for estimating an unknown parameter of a probability distribution: the maximum likelihood estimate (MLE) and the maximum a posteriori probability (MAP) estimate.

Suppose we observe the values of n iid[1] random variables $X_1, \ldots, X_n$ drawn from a single Bernoulli distribution with parameter θ. In other words, for each X_i, we know that

$$P(X_i = 1) = \theta \quad \text{and} \quad P(X_i = 0) = 1 - \theta.$$

Our goal is to estimate the value of θ from these observed values of X_1 through X_n.

[1] iid means Independent, Identically Distributed.

Maximum Likelihood Estimation

The first estimator of θ that we consider is the maximum likelihood estimator. For any hypothetical value $\hat{\theta}$, we can compute the probability of observing the outcome $X_1, \ldots, X_n$ if the true parameter value θ were equal to $\hat{\theta}$. This probability of the observed data is often called the *data likelihood*, and the function $L(\hat{\theta})$ that maps each $\hat{\theta}$ to the corresponding likelihood is called the *likelihood function*. A natural way to estimate the unknown parameter θ is to choose the $\hat{\theta}$ that maximizes the likelihood function. Formally,

$$\hat{\theta}^{\mathrm{MLE}} = \operatorname*{argmax}_{\hat{\theta}} L(\hat{\theta}).$$

(a) [**4 Points**] Write a formula for the likelihood function, $L(\hat{\theta})$. Your function should depend on the random variables $X_1, \ldots, X_n$ and the hypothetical parameter $\hat{\theta}$. Does the likelihood function depend on the order of the random variables?

Solution: Since the X_i are independent, we have

$$\begin{aligned} L(\hat{\theta}) &= P_{\hat{\theta}}(X_1, \ldots, X_n) \\ &= \prod_{i=1}^{n} P_{\hat{\theta}}(X_i) \\ &= \prod_{i=1}^{n} \hat{\theta}^{X_i}(1-\theta)^{1-X_i} \\ &= \hat{\theta}^{\#\{X_i=1\}}(1-\hat{\theta})^{\#\{X_i=0\}}, \end{aligned}$$

where $\#\{\cdot\}$ counts the number of X_i for which the condition in braces holds true. In the third line, we used the trick $X_i = \mathbb{I}\{X_i = 1\}$. The likelihood function does not depend on the order of the data.

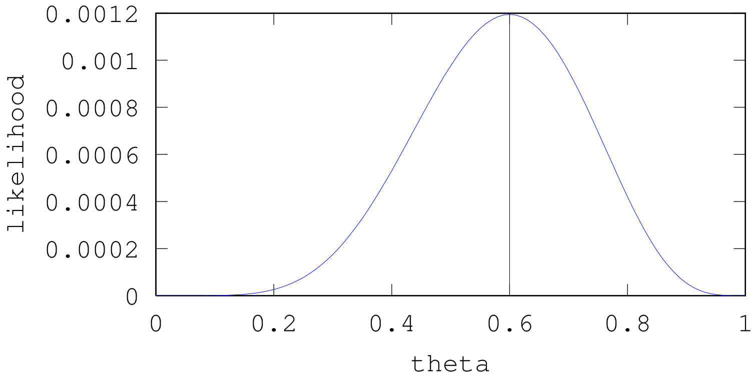
Common mistakes:

- For this problem, many people wrote the definition of the likelihood function $L(\hat{\theta}) = P_{\hat{\theta}}(X_1, \ldots, X_n) = \prod_{i=1}^{n} P_{\hat{\theta}}(X_i)$. Since you knew the distribution of the X_i, we expected you to write out an explicit formula.

(b) [**4 Points**] Suppose that $n = 10$ and the data set contains six 1s and four 0s. Write a short computer program that plots the likelihood function of this data for each value of $\hat{\theta}$ in $\{0, 0.01, 0.02, \ldots, 1.0\}$. For the plot, the x-axis should be $\hat{\theta}$ and the y-axis $L(\hat{\theta})$. Scale your y-axis so that you can see some variation in its value. Please submit both the plot and the code that made it. The plotting code will not be autograded, so please include a printed copy with your solutions.

Solution:

```
thetas = linspace(0,1,101);
likelihoods = thetas.^6 .* (1 .- thetas).^4;
hold("on")
plot(thetas, likelihoods);
plot([0.6, 0.6], [min(likelihoods), max(likelihoods)], "k-")
xlabel("theta");
ylabel("likelihood");
```



Common mistakes:

- For this problem, a few people plotted the log likelihood function

(c) **[4 Points]** Estimate $\hat{\theta}^{\text{MLE}}$ by marking on the x-axis the value of $\hat{\theta}$ that maximizes the likelihood. Find a closed-form formula for the MLE. Does the closed form agree with the plot?

Solution: First, since the log function is increasing, the $\hat{\theta}$ that maximizes the *log likelihood* is the same as the $\hat{\theta}$ that maximizes the likelihood. Let $\ell(\hat{\theta}) = \log(L(\hat{\theta}))$ and introduce the shorthand $n_1 = \#\{X_i = 1\}$ and $n_0 = \#\{X_i = 0\}$. Using the properties of the log function, we can rewrite ℓ as follows

$$\begin{aligned}\ell(\hat{\theta}) &= \log(\theta^{n_1}(1-\theta)^{n_0}) \\ &= n_1 \log(\hat{\theta}) + n_0 \log(1-\hat{\theta})\end{aligned}$$

The first and second derivatives of ℓ are given by

$$\ell'(\hat{\theta}) = \frac{n_1}{\hat{\theta}} - \frac{n_0}{1-\hat{\theta}} \quad \text{and} \quad \ell''(\hat{\theta}) = -\left(\frac{n_1}{\hat{\theta}^2} + \frac{n_0}{(1-\hat{\theta})^2}\right).$$

Since the second derivative is always negative, the function ℓ is concave, and we can find its maximizer by solving $\ell'(\hat{\theta}) = 0$. The solution to this equation is obtained by straight-forward algebra and is given by

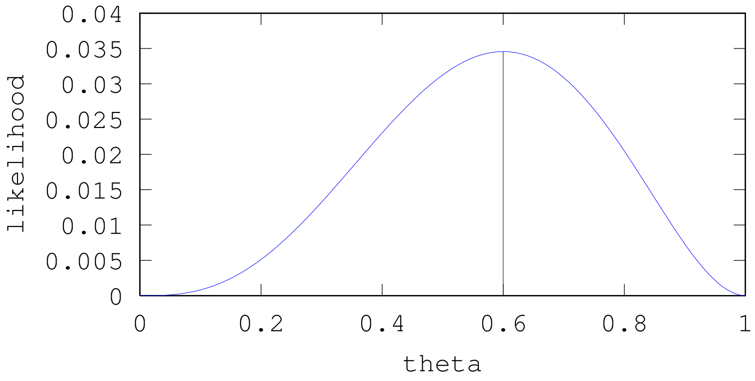
$$\hat{\theta}^{\text{MLE}} = \frac{n_1}{n_1 + n_0}.$$

(d) **[4 Points]** Create three more likelihood plots: one where $n = 5$ and the data set contains three 1s and two 0s; one where $n = 100$ and the data set contains sixty 1s and fourty 0s; and one where $n = 10$ and there are five 1s and five 0s.

Solution:

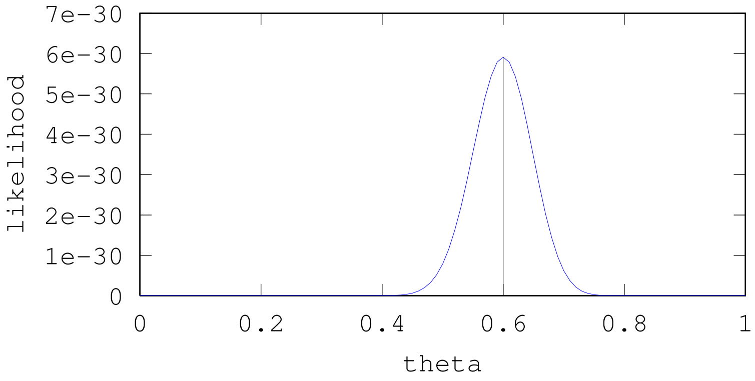
For $n = 5$:

```
thetas = linspace(0,1,101);
likelihoods = thetas.^3 .* (1 .- thetas).^2;
hold("on")
plot(thetas, likelihoods);
plot([0.6, 0.6], [min(likelihoods), max(likelihoods)], "k-")
xlabel("theta");
ylabel("likelihood");
```



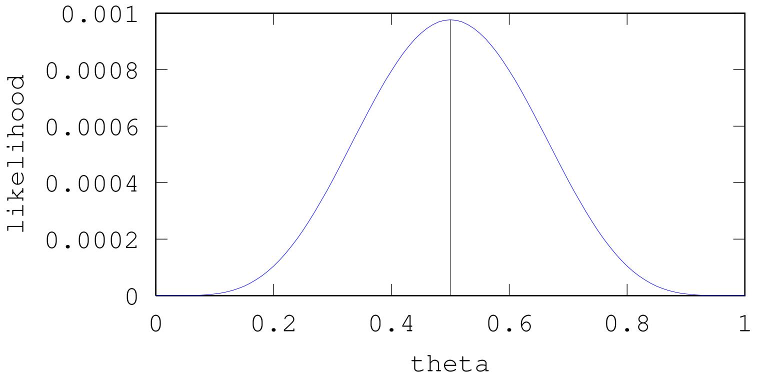
For $n = 100$:

```
thetas = linspace(0,1,101);
likelihoods = thetas.^60 .* (1 .- thetas).^40;
hold("on")
plot(thetas, likelihoods);
plot([0.6, 0.6], [min(likelihoods), max(likelihoods)], "k-")
xlabel("theta");
ylabel("likelihood");
```



For $n = 10$ with five 1s and 0s:

```
thetas = linspace(0,1,101);
likelihoods = thetas.^5 .* (1 .- thetas).^5;
hold("on")
plot(thetas, likelihoods);
plot([0.5, 0.5], [min(likelihoods), max(likelihoods)], "k-")
xlabel("theta");
ylabel("likelihood");
```



(e) [**4 Points**] Describe how the likelihood functions and maximum likelihood estimates compare for the different data sets.

Solution: The MLE is equal to the proportion of 1s observed in the data, so for the first three plots the MLE is always at 0.6 and for the last plot it is at 0.5. As the number of samples n increases, the likelihood function gets more peaked at its maximum value, and the values it takes on decrease.

Common mistakes:

- Many people only compared the MLE estimates, and not the likelihood functions themselves.

Maximum a Posteriori Probability Estimation

In the maximum likelihood estimate, we treated the true parameter value θ as a fixed (non-random) number. In cases where we have some prior knowledge about θ, it is useful to treat θ itself as a random variable, and express our prior knowledge in the form of a prior probability distribution over θ. For example, suppose that the X_1, ..., X_n are generated in the following way:

- First, the value of θ is drawn from a given prior probability distribution
- Second, X_1, ..., X_n are drawn independently from a Bernoulli distribution using this value for θ.

Since both θ and the sequence X_1, ..., X_n are random, they have a joint probability distribution. In this setting, a natural way to estimate the value of θ is to simply choose its most probable value given its prior distribution plus the observed data X_1, ..., X_n.

$$\hat{\theta}^{\text{MAP}} = \underset{\hat{\theta}}{\text{argmax}}\, P(\theta = \hat{\theta}|X_1, \ldots, X_n).$$

This is called the maximum a posteriori probability (MAP) estimate of θ. Using Bayes rule, we can rewrite the posterior probability as follows:

$$P(\theta = \hat{\theta}|X_1, \ldots, X_n) = \frac{P(X_1, \ldots, X_n|\theta = \hat{\theta})P(\theta = \hat{\theta})}{P(X_1, \ldots, X_n)}.$$

Since the probability in the denominator does not depend on $\hat{\theta}$, the MAP estimate is given by

$$\begin{aligned}\hat{\theta}^{\text{MAP}} &= \underset{\hat{\theta}}{\text{argmax}}\, P(X_1, \ldots, X_n|\theta = \hat{\theta})P(\theta = \hat{\theta}) \\ &= \underset{\hat{\theta}}{\text{argmax}}\, L(\hat{\theta})P(\theta = \hat{\theta}).\end{aligned}$$

In words, the MAP estimate for θ is the value $\hat{\theta}$ that maximizes the likelihood function multiplied by the prior distribution on θ. When the prior on θ is a continuous distribution with density function p, then the MAP estimate for θ is given by

$$\hat{\theta}^{\text{MAP}} = \underset{\hat{\theta}}{\text{argmax}}\, L(\hat{\theta})p(\hat{\theta}).$$

For this problem, we will use a Beta(3,3) prior distribution for θ, which has density function given by

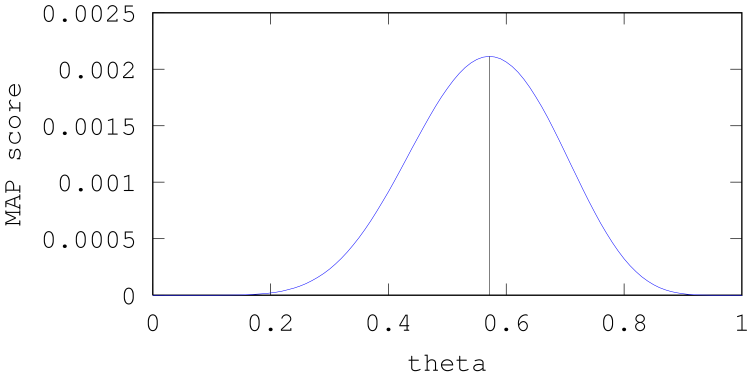
$$p(\hat{\theta}) = \frac{\hat{\theta}^2(1-\hat{\theta})^2}{B(3,3)},$$

where $B(\alpha, \beta)$ is the beta function and $B(3,3) \approx 0.0333$.

(f) **[4 Points]** Suppose, as in part (c), that $n = 10$ and we observed six 1s and four 0s. Write a short computer program that plots the function $\hat{\theta} \mapsto L(\hat{\theta})p(\hat{\theta})$ for the same values of $\hat{\theta}$ as in part (c).

Solution:

```
thetas = linspace(0,1,101);
scores = thetas.^8 .* (1 .− thetas).^6 ./ beta(3,3);
hold(”on”)
plot(thetas, scores);
plot([8/14, 8/14], [min(scores), max(scores)], ”k−”)
xlabel(”theta”);
ylabel(”MAP␣score”);
```



(g) **[4 Points]** Estimate $\hat{\theta}^{\text{MAP}}$ by marking on the x-axis the value of $\hat{\theta}$ that maximizes the function. Find a closed form formula for the MAP estimate. Does the closed form agree with the plot?

Solution: As in the case of the MLE, we will apply the $\log$ function before finding the maximizer. Again, using the notation $n_1 = \#\{X_i = 1\}$ and $n_0 = \#\{X_i = 0\}$, we want to maximize the function

$$\begin{aligned}\ell(\hat{\theta}) &= \log\big(L(\hat{\theta})p(\hat{\theta})\big) \\ &= \log(\hat{\theta}^{n_1+2}(1-\hat{\theta})^{n_0+2}) - \log(B(3,3)).\end{aligned}$$

The normalizing constant for the prior appears as an additive constant and therefore the first and second derivatives are identical to those in the case of the MLE (except with $n_1 + 2$ and $n_0 + 2$ instead of n_1 and n_0, respectively). It follows that the closed form formula for the MAP estimate is given by

$$\hat{\theta}^{\text{MAP}} = \frac{n_1 + 2}{n_1 + n_0 + 4}.$$

Common mistakes:

- A few people used the formula from the slides but forgot to subtract 1 from the numerator and denominator.

(h) **[4 Points]** Compare the MAP estimate to the MLE computed from the same data in part (c). Briefly explain any significant difference.

Solution: The MAP estimate is equal to the MLE with four additional virtual random variables, two that are equal to 1, and two that are equal to 0. This pulls the value of the MAP estimate closer to the value 0.5, which is why the MAP estimate is smaller than the MLE.

(i) **[4 Points]** Comment on the relationship between the MAP and MLE estimates as n goes to infinity.

Solution: As n goes to infinity, the influence of the 4 virtual random variables diminishes, and the two estimators become equal.

Problem 3: Splitting Heuristic for Decision Trees

Recall that the ID3 algorithm iteratively grows a decision tree from the root downwards. On each iteration, the algorithm replaces one leaf node with an internal node that splits the data based on one decision attribute (or feature). In particular, the ID3 algorithm chooses the split that reduces the entropy the most, but there are other choices. For example, since our goal in the end is to have the lowest error, why not instead choose the split that reduces error the most? In this problem we will explore one reason why reducing entropy is a better criterion.

Consider the following simple setting. Let us suppose each example is described by n boolean features: $X = \langle X_1, \dots X_n \rangle$, where $X_i \in \{0, 1\}$, and where $n \geq 4$. Furthermore, the target function to be learned is $f : X \to Y$, where $Y = X_1 \vee X_2 \vee X_3$. That is, $Y = 1$ if $X_1 = 1$ or $X_2 = 1$ or $X_3 = 1$, and $Y = 0$ otherwise. Suppose that your training data contains all of the 2^n possible examples, each labeled by f. For example, when $n = 4$, the data set would be

X_1	X_2	X_3	X_4	Y
0	0	0	0	0
1	0	0	0	1
0	1	0	0	1
1	1	0	0	1
0	0	1	0	1
1	0	1	0	1
0	1	1	0	1
1	1	1	0	1

X_1	X_2	X_3	X_4	Y
0	0	0	1	0
1	0	0	1	1
0	1	0	1	1
1	1	0	1	1
0	0	1	1	1
1	0	1	1	1
0	1	1	1	1
1	1	1	1	1

(a) **[4 Points]** How many mistakes does the best 1-leaf decision tree make, over the 2^n training examples? (The 1-leaf decision tree does not split the data even once)

Solution: A sample X is labeled 0 if and only if $X_1 = X_2 = X_3 = 0$. The number of such binary vectors is given by 2^{n-3} because there are two choices for each of the remaining $n - 3$ features. Since $2^n - 2^{n-3}$ is on the order of 2^n, which is much larger than 2^{n-3}, the best 1-leaf decision tree predicts 1 for every input and makes 2^{n-3} mistakes. This corresponds to making an error $2^{n-3}/2^n = 1/8^{\text{th}}$ of the time.

Common mistakes:

- Many people only answered this question for the specific case when $n = 4$. The question asked about the case when $n \geq 4$.

(b) **[4 Points]** Is there a split that reduces the number of mistakes by at least one? (I.e., is there a decision tree with 1 internal node with fewer mistakes than your answer to part (a)?) Why or why not?

Solution: No. No matter what variable you put at the root, the error rate will remain $1/8$. Splitting on X_i with $i \geq 4$ will split the data so that the proportion of ones in each leaf is $7/8$. In both leaves the tree will predict 1, so it makes the same number of errors as the single-leaf tree that always predicts 1. Splitting on X_1, X_2, or X_3 will split the data into one leaf that is contains only 1s and one leaf where the proportion of 1s is $3/4$. Again, this tree will predict 1 in both leaves, so it makes the same number of mistakes as the single-leaf tree that always predicts 1.

Common mistakes:

- Many people suggested that the tree with a single internal node would make more mistakes than the single-leaf tree that always predicts 1. But, since the tree with an internal node can always predict 1 in both of its leaves, it should never make more mistakes (as long as we choose the predictions in each leaf greedily).

(c) [**4 Points**] What is the entropy of the output label Y for the 1-leaf decision tree (no splits at all)?

Solution: $(1/8)\lg(8) + (7/8)\lg(8/5) = 0.543$

(d) [**4 Points**] Is there a split that reduces the entropy of the output Y by a non-zero amount? If so, what is it, and what is the resulting conditional entropy of Y given this split?

Solution: Ues, splitting with any of x_1 or x_2 or x_3 gives a tree with entropy $(1/2)[0] + 1/2[(1/4)\lg(4) + (3/4)\lg(4/3)] = 0.406$.