# 10-601 Machine Learning: Homework 2

Due 5 p.m. Wednesday, January 28, 2015

## Instructions

- **Late homework policy:** Homework is worth full credit if submitted before the due date, half credit during the next 48 hours, and zero credit after that. You *must* turn in at least $n-1$ of the $n$ homeworks to pass the class, even if for zero credit.
- **Collaboration policy:** Homeworks must be done individually, except where otherwise noted in the assignments. "Individually" means each student must hand in their own answers, and each student must write and use their own code in the programming parts of the assignment. It is acceptable for students to collaborate in figuring out answers and to help each other solve the problems, though you must in the end write up your own solutions individually, and you must list the names of students you discussed this with. We will be assuming that, as participants in a graduate course, you will be taking the responsibility to make sure you personally understand the solution to any work arising from such collaboration.
- **Online submission:** You must submit your solutions online on autolab. We recommend that you use LaTeX, but we will accept scanned solutions as well. On the Homework 2 autolab page, you can download the submission template, which is a tar archive containing blank placeholder pdfs for each of the three problems. Replace each of these pdf files with your solutions for the corresponding problem, create a new tar archive of the top-level directory, and submit your archived solutions online by clicking the "Submit File" button. You should submit a single tar archive identical to the template, except with each of the problem pdfs replaced with your solutions.

  ***DO NOT*** change the name of any of the files or folders in the submission template. In other words, your submitted pdfs should have exactly the same names as those in the submission template. Do not modify the directory structure.

## Problem 1: More Probability Review

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

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

(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.

(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 $X$.

(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 $P(X = 0|X + Y > 0)$.

# 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 $\theta$. 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 $\theta$ from these observed values of $X_1$ through $X_n$.

## Maximum Likelihood Estimation

The first estimator of $\theta$ 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 $\theta$ 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 $\theta$ is to choose the $\hat{\theta}$ that maximizes the likelihood function. Formally,

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

(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?

(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. Please include all plots for this question in the `problem2.pdf` file, as well as the source code for producing them. That is, do not submit the source code and plots as separate files.

(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?

(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.

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

[1] iid means Independent, Identically Distributed.

## Maximum a Posteriori Probability Estimation

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

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

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

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

This is called the maximum a posteriori probability (MAP) estimate of $\theta$. 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

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

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

$$\hat{\theta}^{\text{MAP}} = \operatorname*{argmax}_{\hat{\theta}} L(\hat{\theta}) p(\hat{\theta}).$$

For this problem, we will use a Beta(3,3) prior distribution for $\theta$, 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).

(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?

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

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

## 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: $X = \langle X_1, \ldots X_n \rangle$, 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)

(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?

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

(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?