DS-GA 1003: Machine Learning and Computational Statistics
Homework 7: Bayesian Modeling

Due: Monday, May 9, 2016, at 6pm (Submit via NYU Classes)

Instructions: Your answers to the questions below, including plots and mathematical work, should be submitted as a single file, either HTML or PDF. You may include your code inline or submit it as a separate file. You may either scan hand-written work or, preferably, write your answers using software that typesets mathematics (e.g. LaTeX, LyX, or MathJax via iPython).

1 Introduction

In this homework we work through several basic concepts in Bayesian statistics via one of the simplest problems there is: estimating the probability of heads in a coin flip. Later we'll extend this to the probability of estimating click-through rates in mobile advertising.

2 Coin Flipping: Maximum Likelihood

1. Suppose we flip a coin and get the following seqeunce of heads and tails:

$$\mathcal{D} = (H, H, T)$$

   Give an expression for the probability of observing $\mathcal{D}$ given that the probability of heads is θ. That is, give an expression for $p(\mathcal{D} \mid \theta)$. This is called the **likelihood of θ for the data $\mathcal{D}$**.

2. How many different sequences of 3 coin tosses have 2 heads and 1 tail? If we coss the coin 3 times, what is the probability of 2 heads and 1 tail? (Answer should be in terms of θ.)

3. More generally, give an expression for the likelihood $p(\mathcal{D} \mid \theta)$ for a particular sequence of flips $\mathcal{D}$ that has n_h heads and n_t tails. Make sure you have expressions that make sense even for $\theta = 0$ and $n_h = 0$, and other boundary cases. You may use the convention that $0^0 = 1$, or you can break your expression into cases if needed.

4. Prove that the maximum likelihood estimate of θ given we observed a sequence with n_h heads and n_t tails is

$$\hat{\theta}_{\text{MLE}} = \frac{n_h}{n_h + n_t}.$$

(Hint: Maximizing the log-likelihood is equivalent and is often easier. As usual, make sure everything make sense for the boundary cases, such as data with only heads.)

3 Coin Flipping: Bayesian Approach with Beta Prior

We'll now take a Bayesian approach to the coin flipping problem, in which we treat θ as a random variable sampled from some prior distribution $p(\theta)$. We'll represent the ith coin flip by a random variable $X_i \in \{0,1\}$, where $X_i = 1$ if the ith flip is heads. We assume that the X_i's are conditionally indendent given θ. This means that the joint distribution of the coin flips and θ factorizes as follows:

$$\begin{aligned} p(x_1,\ldots,x_n,\theta) &= p(\theta)p(x_1,\ldots,x_n \mid \theta) \text{ (always true)} \\ &= p(\theta)\prod_{i=1}^{n} p(x_i \mid \theta) \text{ (by conditional independence).} \end{aligned}$$

1. Suppose that our prior distribution on θ is $\text{Beta}(h,t)$, for some $h,t > 0$. That is, $p(\theta) \propto \theta^{h-1}(1-\theta)^{t-1}$. Suppose that our sequence of flips $\mathcal{D}$ has n_h heads and n_t tails. Show that the posterior distribution for θ is $\text{Beta}(h+n_h, t+n_t)$. That is, show that

   $$p(\theta \mid \mathcal{D}) \propto \theta^{h-1+n_h}(1-\theta)^{t-1+n_t}.$$

   We say that the Beta distribution is **conjugate** to the Bernoulli distribution since the prior and the posterior are both in the same family of distributions (i.e. both Beta distributions).

2. Give expressions for the MLE, the MAP, and the posterior mean estimates of θ. [Hint: You may use the fact that a $\text{Beta}(h,t)$ distribution has mean $h/(h+t)$ and has mode $(h-1)/(h+t-2)$ for $h,t > 1$. For the Bayesian solutions, you should note that as $h+t$ gets very large, the posterior mean and MAP approach the prior mean $h/(n+h)$, while when for]

3. What happens to $\hat{\theta}_{\text{MLE}}$, $\hat{\theta}_{\text{MAP}}$, and $\hat{\theta}_{\text{POSTERIOR MEAN}}$ as the number of coin flips approaches infinity?

4. Which of the MLE, MAP, and posterior mean estimates are **unbiased** estimates of θ, if any? [Hint: The answer may depend on the parameters h and t of the prior.]

5. Suppose somebody gives you a coin and asks you to give an estimate of the probability of heads, but you can only toss the coin 3 times. You have no particular reason to believe this is an unfair coin. Would you prefer the MLE or the posterior mean as a point estimate of θ? If the posterior mean, what would you use for your prior?

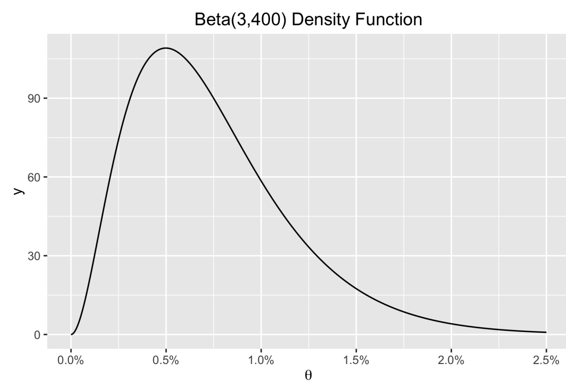
4 Hierarchical Bayes for Click-Through Rate Estimation

In mobile advertising, ads are often displayed inside apps on a phone or tablet device. When an ad is displayed, this is called an "impression." If the user clicks on the ad, that is called a "click." The probability that an impression leads to a click is called the "click-through rate" (CTR).

Suppose we have $d = 1000$ apps. For various reasons, each app tends to have a different overall CTR. For the purposes of designing an ad campaign, we want estimates of all the app-level CTRs, which we'll denote by $\theta_1, \ldots, \theta_d$. Of course, the particular user seeing the impression and the particular ad that is shown have an effect on the CTR, but we'll ignore these issues for now. [Because so many clicks on mobile ads are accidental, it turns out that the overall app-level CTR often dominates the effect of the particular user and the specific ad.]

If we have enough impressions for a particular app, then the empirical fraction of clicks will give a good estimate for the actual CTR. However, if we have relatively few impressions, we'll have some problems using the empirical fraction. Typical CTRs are less than 1%, so it takes a fairly large number of observations to get a good estimate of CTR. For example, even with 100 impressions, the only possible CTR estimates are $0\%, 1\%, 2\%, \ldots, 100\%$. The 0% estimate is almost certainly much too low, and anything 2% or higher is almost certainly much too high. Our goal is to come up with reasonable point estimates for $\theta_1, \ldots, \theta_{1000}$, even when we have very few observations for some apps.

If we wanted to apply the Bayesian approach worked out in the previous problem, we could come up with a prior that seemed reasonable. For example, we could use the following $\text{Beta}(3, 400)$ as a prior distribution on each θ_i:



In this basic Bayesian approach, the parameters 3 and 400 would be chosen by the data scientist based on prior experience, or "best guess", but without looking at the new data. Another approach would be to use the data to help you choose the parameters a and b in $\text{Beta}(a, b)$. This would **not** be a Bayesian approach, though it is frequently used in practice. One method in this direction is called **empirical Bayes**. Empirical Bayes can be considered a frequentist approach, in which estimate a and b from the data $\mathcal{D}$ using some estimation technique, such as maximum likelihood. The proper Bayesian approach to this type of thing is called **hierarchical Bayes**, in which we put another prior distribution on a and b.

Mathematical Description

We'll now give a mathematical description of our model, assuming the prior parameters a and b are given. Let $n_1, \ldots, n_d$ be the number of impressions we observe for each of the d apps. In this problem, we will not consider these to be random numbers. For the ith app, let $c_i^1, \ldots, c_i^{n_i} \in \{0, 1\}$ be indicator variables determining whether or not each impression was clicked. That is, $c_i^j = 1(j\text{th impression on } i\text{th app was clicked})$. We can summarize the data on the ith app by $\mathcal{D}_i = (x_i, n_i)$, where $x_i = \sum_{j=1}^{n_i} c_i^j$ is the total number of impressions that were clicked for app i. Let $\theta = (\theta_1, \ldots, \theta_d)$, where θ_i is the CTR for app i.

In our Bayesian approach, we act as though the data were generated as follows:

1. Sample $\theta_1, \ldots, \theta_d$ i.i.d. from $\text{Beta}(a, b)$.
2. For each app i, sample $c_i^1, \ldots, c_i^{n_i}$ i.i.d. from $\text{Bernoulli}(\theta_i)$.

4.1 Empirical Bayes for a single app

We start by working out some details for Bayesian inference for a single app. That is, suppose we only have the data $\mathcal{D}_i$ from app i, and nothing else. Mathematically, this is exactly the same setting as the coin tossing setting above, but here we push it further.

1. Give an expression for $p(\mathcal{D}_i \mid \theta_i)$, the likelihood of $\mathcal{D}_i$ given the probability of click θ_i, in terms of θ_i, x_i and n_i.

2. The probability density for the $\text{Beta}(a, b)$ distribution, evaluated at θ_i, is given by

$$\text{Beta}(\theta_i; a, b) = \frac{1}{B(a,b)} \theta_i^{a-1} \left(1 - \theta_i\right)^{b-1}$$

where $B(a, b)$ is called the Beta function. Explain why we must have

$$\int \theta_i^{a-1} \left(1 - \theta_i\right)^{b-1} \, d\theta = B(a, b),$$

and give the full density function for the prior on θ_i, in terms of a, b, and the normalization function B.

3. Give an expression for the posterior distribution $p(\theta_i \mid \mathcal{D}_i)$. In this case, include the constant of proportionality. (In other words, do not use the "is proportional to" sign $\propto$ in your final expression.) [Hint: This problem is essentially a repetition of an earlier problem.]

4. Give a closed form expression for $p(\mathcal{D}_i)$, the marginal likelihood of $\mathcal{D}_i$, in terms of the a, b, x_i, and n_i. You may use the normalization function $B(\cdot, \cdot)$ for convenience, but you should not have any integrals in your solution. (Hint: $p(\mathcal{D}_i) = \int p\left(\mathcal{D}_i \mid \theta_i\right) p(\theta_i) \, d\theta_i$, and the answer will be a ratio of two beta function evaluations.)

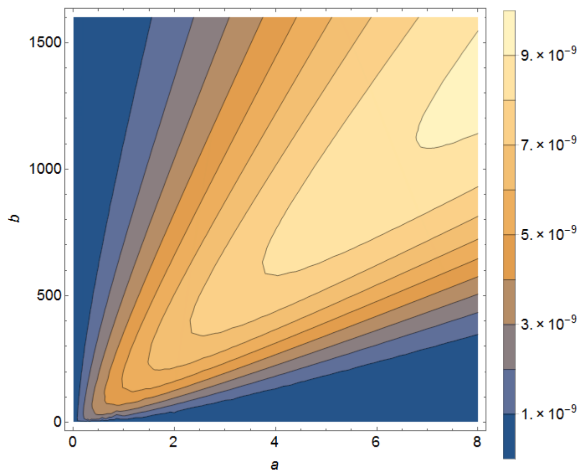


Figure 1: A plot of $p(\mathcal{D}_i \mid a, b)$ as a function of a and b.

5. The maximum likelihood estimate for θ_i is x_i/n_i. Let $p_{\text{MLE}}(\mathcal{D}_i)$ be the marginal likelihood of $\mathcal{D}_i$ when we use a prior on θ_i that puts all of its probability mass at x_i/n_i. Note that

$$\begin{aligned} p_{\text{MLE}}(\mathcal{D}_i) &= p\left(\mathcal{D}_i \mid \theta_i = \frac{x_i}{n_i}\right) p\left(\theta_i = \frac{x_i}{n_i}\right) \\ &= p\left(\mathcal{D}_i \mid \theta_i = \frac{x_i}{n_i}\right). \end{aligned}$$

Explain why, or prove, that $p_{\text{MLE}}(\mathcal{D}_i)$ is larger than $p(\mathcal{D}_i)$ for any other prior we might put on θ_i. If it's too hard to reason about all possible priors, it's fine to just consider all Beta priors. [Hint: This does not require much or any calculation. It may help to think about the integral $p(\mathcal{D}_i) = \int p(\mathcal{D}_i \mid \theta_i)\, p(\theta_i)\, d\theta_i$ as a weighted average of $p(\mathcal{D}_i \mid \theta_i)$ for different values of θ_i, where the weights are $p(\theta_i)$.]

6. One approach to getting an **empirical Bayes** estimate of the parameters a and b is to use maximum likelihood. Such an empirical Bayes estimate is often called an **ML-2** estimate, since it's maximum likelihood, but at a higher level in the Bayesian hierarchy. To emphasize the dependence of the likelihood of $\mathcal{D}_i$ on the parameters a and b, we'll now write it as $p(\mathcal{D}_i \mid a, b)$[1]. The empirical Bayes estimates for a and b are given by

$$(\hat{a}, \hat{b}) = \underset{(a,b) \in \mathbf{R}^{>0} \times \mathbf{R}^{>0}}{\arg\max}\; p(\mathcal{D}_i \mid a, b).$$

To make things concrete, suppose we observed $x_i = 3$ clicks out of $n_i = 500$ impressions. A plot of $p(\mathcal{D}_i \mid a, b)$ as a function of a and b is given in Figure 1. It appears from this plot that

[1]Note that this is a slight (though common) abuse of notation, because a and b are not random variables in this setting. It might be more appropriate to write this as $p(\mathcal{D}_i; a, b)$ or $p_{a,b}(\mathcal{D}_i)$. But this isn't very common.

the likelihood will keep increasing as a and b increase, at least if a and b maintain a particular ratio. Indeed, this likelihood function never attains its maximum, so we cannot use ML-2 here. Explain what's happening to the prior as we continue to increase the likelihood. [Hint: It is a property of the Beta distribution (not difficult to see), that for any $\theta \in (0, 1)$, there is a Beta distribution with expected value θ and variance less than ε, for any $\varepsilon > 0$. What's going in here is similar to what happens when you attempt to fit a gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ to a single data point using maximum likelihood.]

4.2 Empirical Bayes Using All App Data

In the previous section, we considered working with data from a single app. With a fixed prior, such as Beta(3,400), our Bayesian estimates for θ_i seem more reasonable (to me, the person who chose the prior) than the MLE when our sample size n_i is small. The fact that these estimates seem reasonable is an immediate consequence of the fact that I chose the prior to give high probability to estimates that seem reasonable to me, before ever seeing the data. Our attempt to use empirical Bayes (ML-2) to choose the prior in a data-driven way was not successful. With only a single app, we were essentially overfitting the prior to the data we have. In this section, we'll consider using the data from all the apps, in which case empirical Bayes makes more sense.

1. Let $\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_d)$ be the data from all the apps. Give an expression for $p(\mathcal{D} \mid a, b)$, the **marginal likelihood** of $\mathcal{D}$. Expression should be in terms of a, b, x_i, n_i for $i = 1, \ldots, d$. (Hint: This problem should be easy, based on a problem from the previous section.)