Homework 1: parameter estimation

CSCI145/MATH166, Mike Izbicki

DUE: Thursday, 26 September at the beginning of class

Name:

If you complete this assignment in LaTeX, then you will receive +2 pts extra credit.

Problem 1. (10pts) In class, we computed the maximum likelihood estimator for the mean of a normal distribution. For this problem, you will compute the maximum likelihood estimator for the standard deviation.

Let $X = \{x_1, ..., x_n\}$ be a set of i.i.d. normally distributed random variables with mean μ and standard deviation σ. Recall that the maximum likelihood estimator of μ and σ is the solution to

$$\hat{\mu}_{mle}, \hat{\sigma}_{mle} = \arg\max_{\mu,\sigma} p(X|\mu, \sigma^2). \tag{1}$$

Derive a formula for $\hat{\sigma}_{mle}$. In your derivation, you may use the fact that $\hat{\mu}_{mle} = \frac{1}{n}\sum_{i=1}^{n} x_i$ that we proved in class.

Problem 2. (10pts) The poisson distribution has density

$$p(x|\lambda) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}, \tag{2}$$

where x is a non-negative integer. Let $X = \{x_1, ..., x_n\}$ be a set of i.i.d. poisson random variables, and calculate the maximum likelihood estimate of λ.

Problem 3. (10pts) The exponential distribution has density

$$p(x|\lambda) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

Let $X = \{x_1, ..., x_n\}$ be a set of i.i.d. exponentially distributed random variables, and calculate the maximum likelihood estimate of λ.

HINT: The following problems are a bit different than the previous problems. The normal, poisson, and exponential distributions are all members of the so-called *exponential family* of distributions. This family of distributions is called the *exponential* family because the exponential function appears in each of their densities in a special way. Section 2.4 of Bishop provides details, but the most important property of these distributions is that their maximum likelihood parameter estimates can be entirely characterized by the arithmetic mean of the samples. (Hopefully you saw that pattern in your solutions to the previous problems.)

The laplace and power distributions, however, are not a member of the exponential family. (The densities contain exponentials, but not in the right way.) The maximum likelihood parameter estimate is therefore not related to the arithmetic mean of the samples. Instead, it is related to one of the following "generalized means": either the geometric mean, harmonic mean, median, or mode. Your solution two the next two problems should be related to one of these "generalized means," and knowing this fact may help perform some of the calculations.

Problem 4. (10pts) The power distribution has density

$$p(x|\theta) = \theta x^{\theta-1}. \tag{4}$$

Calculate the maximum likelihood estimator of θ.

Problem 5. (20pts) The laplace distribution has density

$$p(x|\mu, b) = \frac{1}{b} \exp\left(- \frac{|x - \mu|}{b} \right). \tag{5}$$

Calculate the maximum likelihood estimator of μ and b.

Problem 6. (25pts) We now switch gears to consider *maximum a posteriori* (MAP) estimation.

Let $X = \{x_1, ..., x_n\}$ be i.i.d. normally distributed random variables with mean μ and variance 1, and let the prior distribution over μ be $\mathcal{N}(\mu|\mu_0, \sigma_0)$. Recall that the MAP estimate of μ is

$$\hat{\mu}_{map} = \arg\max_{\mu} p(\mu|X). \tag{6}$$

(a) Derive a formula for $\hat{\mu}_{map}$.

(b) Calculate $\lim_{\sigma_0 \to \infty} \hat{\mu}_{map}$.

(c) Calculate $\lim_{\sigma_0 \to 0} \hat{\mu}_{map}$.

Problem 7. (15pts) The purpose of this problem is to begin connecting the theoretical frameworks we're developing to a real world application.

We have a collection Y of newspapers, and a collection T of topics. For example, the newspapers might be the *Scripps Voice*, the *CMC Forum*, the *Student Life*, and the *Claremont Independent*; and the topics might be "women," "stags," "academics," and "politics." The goal of our analysis is to see which newspapers are most likely to publish stories about which topics.

Formally, our goal is to calculate

$$\arg\max_y p(Y = y|T = t) \tag{7}$$

for any topic t. For example, we might expect that all four newspapers will write about the topic of "women" occasionally, but that the *Scripps Voice* will write about women the most. If this is true, we would have

$$\arg\max_y p(Y = y|T = \text{women}) = \textit{Scripps Voice}. \tag{8}$$

In order to compute (7) for a particular topic, we need to make assumptions about the form of the probability distribution $P(Y|T)$. One possible assumption is that $P(Y|T)$ has what Bishop calls a multinomial distribution. (Other books call this a categorical distribution.) Under this assumption, we have that for each $y \in Y$ and $t \in T$,

$$p(Y = y|T = t, \mu) = \mu_{y,t}, \tag{9}$$

where μ is a $|Y| \times |T|$ dimensional matrix satisfying $\mu_{y,t} \geq 0$ for all y and t, and $\sum_{y \in Y}^{m} \mu_{y,t} = 1$ for all t. **Note:** Equation (9) above is equivalent to Equation (2.26) in Bishop, and to complete the rest of this problem you will need to understand why.

(a) We are given a set of n documents. Each document i has been labeled with its topic t_i and its newspaper y_i, and we assume these values are sampled i.i.d. from $P(Y|T, \mu)$ defined above. What is the maximum likelihood estimator for μ?

Solving for this maximum likelihood estimator requires a relatively advanced technique called Lagrange multipliers. You may lookup the solution in Bishop section 2.2 and simply report the answer here.

(b) Suppose we have a dirichlet prior over μ. Then what is the MAP estimate for μ? As before, this solution requires Lagrange multipliers, and so you may simply report the solution from Bishop here.

(c) There are several other ways we could model the distribution $p(Y|T)$. One other possibility is to use Bayes theorem to show that $p(Y|T) \propto p(T|Y)p(Y)$, and model both $p(T|Y)$ and $p(Y)$ as multinomial distributions. These are two models are closely related, but can give different answers to the solution of (7). Which of these two models makes the most sense for this problem, and why?