Statistical Machine Learning GU4241/GR5241
Spring 2017
https://courseworks.columbia.edu/

Linxi Liu
ll3098@columbia.edu
Shuaiwen Wang
sw2853@columbia.edu
Haolei Weng
hw2375@columbia.edu

Homework 1

Due: Tuesday, Feb. 7th, 2017

Homework submission: We will collect your homework **at the beginning of class** on the due date. You also need to submit **your code** for Problem 2 and Problem 3 through Canvas. If you cannot attend class that day, you can leave your solution in the homework dropbox 904 at the Department of Statistics, 9th floor SSW, at any time before then.

Problem 1 (Maximum Likelihood Estimation, 10 points)

In this problem, we analytically derive maximum likelihood estimators for the parameters of an example model distribution, the gamma distribution.

The gamma distribution is univariate (one-dimensional) and continuous. It is controlled by two parameters, the *location parameter* μ and the *shape parameter* ν. For a gamma-distributed random variable X, we write $X \sim \mathcal{G}(\mu, \nu)$. $\mathcal{G}$ is defined by the following density function:

$$p(x|\mu,\nu) := \left(\frac{\nu}{\mu}\right)^{\nu} \frac{x^{\nu-1}}{\Gamma(\nu)} \exp\left(-\frac{\nu x}{\mu}\right) ,$$

where $x \geq 0$ and $\mu, \nu > 0$.[1] Whenever $\nu > 1$, the gamma density has a single peak, much like a Gaussian. Unlike the Gaussian, it is not symmetric. The first two moment statistics of the gamma distribution are given by

$$\mathsf{E}[X] = \mu \qquad \text{and} \qquad \mathsf{Var}[X] = \frac{\mu^2}{\nu} \tag{1}$$

for $X \sim \mathcal{G}(\mu, \nu)$. The plots in Figure 1 should give you a rough idea of what the gamma density may look like and how different parameter values influence its behavior.

Homework questions:

1. Write the general analytic procedure to obtain the maximum likelihood estimator (including logarithmic transformation) in the form of a short algorithm or recipe. A few words are enough, but be precise: Write all important mathematical operations as formulae. Assume that data is given as an i. i. d. sample $x_1, \ldots, x_n$. Denote the conditional density in question by $p(x|\theta)$, and the likelihood by $l(\theta)$. Make sure both symbols show up somewhere in your list, as well as a logarithm turning a product into a sum.

2. Derive the ML estimator for the location parameter μ, given data values $x_1, \ldots, x_n$. Conventionally, an estimator for a parameter is denoted by adding a hat: $\hat{\mu}$. Considering the expressions in (1) for the mean and variance of the gamma distribution, and what you know about MLE for Gaussians, the result should not come as a surprise.

3. A quick look at the gamma density will tell you that things get more complicated for the shape parameter: ν appears inside the gamma function, and both inside and outside the exponential. Thus, instead of deriving

[1] The symbol Γ denotes the distribution's namesake, the *gamma function*, defined by

$$\Gamma(\nu) := \int_0^\infty e^{-t} t^{\nu-1} dt .$$

The gamma function is a generalization of the factorial to the real line: $\Gamma(n) = (n-1)!$ for all $n \in \mathbf{N}$. Fortunately, we will not have to make explicit use of the integral.

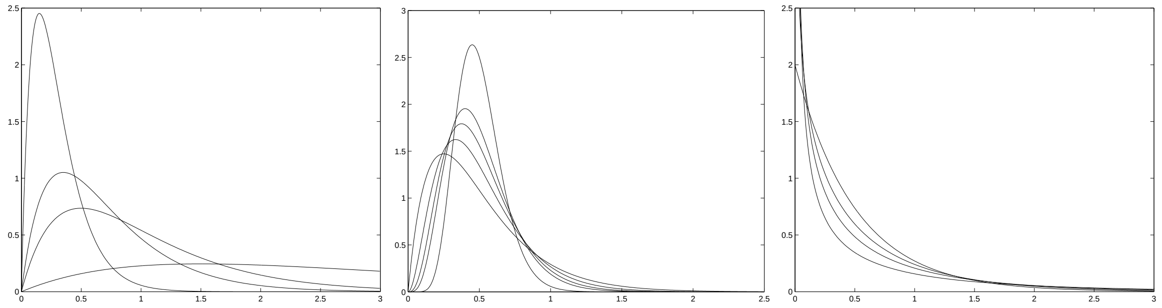

Figure 1: *Left:* The plot shows the density for different values of the location parameter ($\mu = 0.3, 0.5, 1.0, 3.0$), with the shape parameter fixed to $\nu = 2$. Since $\nu > 1$, the densities peak. As we increase μ, the peak moves to the right, and the curve flattens. *Middle:* For $\mu = 0.5$ fixed, we look at different values of the shape parameter ($\nu = 2, 3, 4, 5, 19$). Again, all the densities peak, with the peak shifting to the right as we increase ν. *Right:* If $\nu < 1$, the density turns into a monotonously decreasing function. The smaller the value of ν, the sharper the curve dips towards the origin.

a formula of the form $\hat{\nu} := \ldots$, please show the following: Given an i. i. d. data sample $x_1, \ldots, x_n$ and the value of μ, the ML estimator $\hat{\nu}$ for the gamma distribution shape parameter solves the equation

$$\sum_{i=1}^{n} \left(\ln \left(\frac{x_i \hat{\nu}}{\mu} \right) - \left(\frac{x_i}{\mu} - 1 \right) - \phi\left(\hat{\nu}\right) \right) = 0 \, .$$

The symbol ϕ is a shorthand notation for

$$\phi\left(\nu\right) := \frac{\frac{\partial \Gamma(\nu)}{\partial \nu}}{\Gamma\left(\nu\right)} \, .$$

In mathematics, ϕ is known as the *digamma function*.

Problem 2 (Subset Selection Methods, 15 points)

In this problem, we will compare different subset selection methods. We will study the Credit data set, which can be downloaded from CourseWorks. The data set records balance (average credit card debt) as well as several quantitative predictors: age, cards (number of credit cards), education (years of education), income (in thousands of dollars), limit (credit limit), and rating (credit rating). There are also four qualitative variables: gender, student (student status), status (marital status), and ethnicity (Caucasian, African American or Asian). We want to fit a regression model of balance on the rest of the variables.

- *(Best subset selection)* The regsubsets() function in R (part of the leaps library) performs the best subset selection by identifying the best model that contains a given number of predictors, where *best* is defined to be the one which minimizes the residual sum-of-squares (RSS).

  Here we need to represent the qualitative predictors by dummy variables. gender, student and ethnicity are all two-level categorical variables, and each of them is coded by one dummy variable. ethnicity takes on tree values and is coded by two dummy variables. Therefore, we have 11 predictors in total.

- *(Forward stepwise selection)* We can also use the regsubsets() function to perform forward stepwise selection, using the argument method=‘‘forward’’.

- *(Backward stepwise selection)* The regsubsets() function can be used to perform backward stepwise selection as well (method=‘‘backward’’). Here we start from the full model and at each step remove a predictor which leaves a model having smallest RSS.

- *(Choosing the optimal model)* After obtaining a set of models by using the subset selection approaches, we will choose a single best model which minimizes the prediction error. For this problem, we use C_p or BIC statistic as estimates of the prediction error (We will talk about this later in class). C_p statistic is defined by

  $$C_p = \frac{1}{n}(\text{RSS} + 2p\hat{\sigma}^2),$$

  where $\hat{\sigma}^2$ is an estimate of the variance of the error. BIC is defined by

  $$\text{BIC} = \frac{1}{n}(\text{RSS} + \log(n)p\hat{\sigma}^2).$$

  The summary() function returns RSS, C_p and BIC. You DO NOT need to compute them by yourselves.

Homework problems.

1. Apply the three subset selection methods mentioned above to Credit data set. Plot the RSS as a function of the number of variables for these three methods in the same figure.

2. Each subset selection method results in a set of models. For each approach, choose a single optimal model by using C_p and BIC statistics respectively. Report the optimal models for each approach (i.e. specify the predictors in the optimal model).

Remark. From this problem, you may notice that BIC tends to select a model with less predictors when compared to C_p.

Problem 3 (PCA, 15 points)

1. For each of the 30 stocks in the Dow Jones Industrial Average, download the closing prices for every trading day from January 1, 2010 to January 1, 2011. You can use http://finance.yahoo.com to find the data. To download the prices for symbol AA, for example, use the url http://ichart.finance.yahoo.com/table.csv?s=AA&a=00&b=1&c=2010&d=00&e=1&f=2011&g=d&ignore=.csv This is the URL generated by Yahoo finance when we try to download their historical prices to a spreadsheet. See this page for the page in which the above link can be found. Please find a way to download the data efficiently.

2. Perform a PCA on the prices and create the biplot (call function `princomp()` and use `cor=FALSE`). Do you see any structure in the biplot, perhaps in terms of the types of stocks? How about the screeplot – how many important components seem to be in the data?

3. Repeat part 2 with `cor=TRUE`. This is equivalent to scale each column of the data matrix.

4. Calculate the return for each stock, and repeat part 3 on the return data. In looking at the screeplot, what does this tell you about the 30 stocks in the DJIA? If each stock were fluctuating up and down randomly and independent of all the other stocks, what would you expect the screeplot to look like?