10-601 Machine Learning, Fall 2011: Homework 2

Machine Learning Department
Carnegie Mellon University

Due: October 6th, 2011, 5pm

Instructions There are 2 questions on this assignment. The second question involves coding, so start early. Please submit your completed homework to Sharon Cavlovich (GHC 8215) by 5 PM on Thursday, October 6th, 2011. Submit your homework as 2 **separate** sets of pages, one for each question. Please staple or paperclip all pages from a single question together, but **DO NOT** staple pages from different questions together. This will make it much easier for the TA's to split up your homework for grading. Include your name and email address on each set.

1 MLE and MAP [William Bishop, 20 points]

In this problem we will find the maximum likelihood estimator (MLE) and maximum a posteriori (MAP) estimator for the mean of a univariate normal distribution. Specifically, we assume we have N samples, $x_1, \ldots, x_N$ independently drawn from a normal distribution with *known* variance σ^2 and *unknown* mean μ.

1. [5 Points] Please derive the MLE estimator for the mean μ. Make sure to show all of your work.
2. [12 Points] Now derive the MAP estimator for the mean μ. Assume that the prior distribution for the mean is itself a normal distribution with mean ν and variance β^2. Please show all of your work. HINT: You may want to make use of the fact that:

$$\beta^2 \left(\sum_{i=1}^{N} (x_i - \mu)^2 \right) + \sigma^2(\mu - \nu)^2 =$$

$$\left[\mu\sqrt{N\beta^2 + \sigma^2} - \frac{\sigma^2\nu + \beta^2 \sum_{i=1}^{N} x_i}{\sqrt{N\beta^2 + \sigma^2}} \right]^2 - \frac{[\sigma^2\nu + \beta^2 \sum_{i=1}^{N} x_i]^2}{N\beta^2 + \sigma^2} + \beta^2 \left(\sum_{i=1}^{N} x_i^2 \right) + \sigma^2\nu^2$$

3. [3 Points] Please comment on what happens to the MLE and MAP estimators as the number of samples N goes to infinity.

1.1 Answer to Part A

First, we derive the likelihood term:

$$\begin{aligned} P(x_1, \ldots, x_N|\mu) &= \prod_{i=1}^{N} P(x_i|\mu) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \end{aligned}$$

Next, we note that log is a monotonically increase function, so we can maximize the log-likelihood:

$$\log\left(P(x_1,\ldots,x_N|\mu)\right) = \sum_{i=1}^{N} \log(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{(x_i-\mu)^2}{2\sigma^2}$$

We take derivatives of this with respect to μ and find:

$$\frac{d\log\left(P(x_1,\ldots,x_N|\mu)\right)}{d\mu} = \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2}$$

Setting the left hand side equal to zero, we find:

$$\begin{aligned}
0 &= \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} \\
0 &= \sum_{i=1}^{N} (x_i-\mu) \\
\sum_{i=1}^{N} \mu &= \sum_{i=1}^{N} x_i \\
N\mu &= \sum_{i=1}^{N} x_i \\
\hat{\mu} &= \frac{\sum_{i=1}^{N} x_i}{N}
\end{aligned}$$

1.2 Answer to Part B

~~There are two ways to solve this problem~~. We first show an easier way, which is sufficient to find the MAP estimator.

We use Bayes' rule to write:

$$P(\mu|x_1,\ldots x_n) = \frac{P(x_1,\ldots,x_N|\mu)P(\mu)}{P(x_1,\ldots,x_N)}$$

From part A, we know that we can write:

$$P(x_1,\ldots,x_N|\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

and we are given that:

$$P(\mu) = \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(\mu-\nu)^2}{2\beta^2}}$$

Thus, we desire to find the value of μ which maximizes:

$$P(\mu|x_1,\ldots x_n) = \frac{\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(\mu-\nu)^2}{2\beta^2}}}{C}$$

where $C = P(x_1, \ldots, x_N)$.

We note that because we are simply looking for the value of μ that maximizes this expression we can take the log of both side and write:

$$\log\left(P(\mu|x_1, \ldots x_n)\right) = \left(\sum_{i=1}^{N} -\log\left(\sqrt{2\pi\sigma^2}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right) - \log\left(\sqrt{2\pi\beta^2}\right) - \frac{(\mu-\nu)^2}{2\beta^2}$$

Taking the derivative with respect to μ, we have:

$$\frac{\partial \log\left(P(\mu|x_1, \ldots x_n)\right)}{\partial\mu} = \left(\sum_{i=1}^{N} \frac{x_i-\mu}{\sigma^2}\right) - \frac{\mu-\nu}{\beta^2}$$

Setting this equal to zero, we have:

$$\begin{aligned}
0 &= \left(\sum_{i=1}^{N} \frac{x_i-\mu}{\sigma^2}\right) - \frac{\mu-\nu}{\beta^2} \\
\frac{\mu-\nu}{\beta^2} &= \sum_{i=1}^{N} \frac{x_i-\mu}{\sigma^2} \\
\frac{\mu-\nu}{\beta^2} &= -\frac{\sum_{i=1}^{N} x_i}{\sigma^2} - \frac{N\mu}{\sigma^2} \\
\frac{\mu}{\beta^2} + \frac{N\mu}{\sigma^2} &= \frac{\sum_{i=1}^{N} x_i}{\sigma^2} + \frac{\nu}{\beta^2} \\
\frac{(\sigma^2 + N\beta^2)\mu}{\sigma^2\beta^2} &= \frac{\sigma^2\nu + \beta^2\sum_{i=1}^{N} x_i}{\sigma^2\beta^2} \\
\hat{\mu} &= \frac{\sigma^2\nu + \beta^2\sum_{i=1}^{N} x_i}{\sigma^2 + N\beta^2}
\end{aligned}$$

~~The second way involves a little more work. In this way, we first show that the posterior distribution is itself a Gaussian, and we then use the fact that the mean of a Gaussian is where it achieves its maximum value to find the MAP.~~

~~We start with the fact that we again want to find:~~

$$P(\mu|x_1, \ldots x_n) = \frac{\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{\sigma^2}}\right) \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(\mu-\nu)^2}{2\beta^2}}}{C}$$

~~where $C = P(x_1, \ldots, x_N)$. However, instead of simply maximizing this function, we first show that $P(\mu|x_1, \ldots x_n)$ is itself a Gaussian. Note, that we can write:~~

$$\begin{aligned}
P(\mu|x_1, \ldots x_n) &\propto \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(\mu-\nu)^2}{2\beta^2}} \\
&\propto \left(\prod_{i=1}^{N} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) e^{-\frac{(\mu-\nu)^2}{2\beta^2}} \\
&= e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}(x_i-\mu)^2\right)} e^{-\frac{(\mu-\nu)^2}{2\beta^2}} \\
&= e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}(x_i-\mu)^2\right) - \frac{(\mu-\nu)^2}{2\beta^2}} \\
&= e^{-\frac{\beta^2\left(\sum_{i=1}^{N}(x_i-\mu)^2\right) + \sigma^2(\mu-\nu)^2}{2\sigma^2\beta^2}}
\end{aligned}$$