Homework 4 Solution

October 31, 2017

Due: October 17, 2017, 11:59 PM EST

Instructions

Your homework submission must cite any references used (including articles, books, code, websites, and personal communications). All solutions must be written in your own words, and you must program the algorithms yourself. If you do work with others, you must list the people you worked with. Submit your solutions as a PDF to the E-Learning at UF (http://elearning.ufl.edu/).

Your programs must be written in either MATLAB or Python. The relevant code to the problem should be in the PDF you turn in. If a problem involves programming, then the code should be shown as part of the solution to that problem. If you solve any problems by hand just digitize that page and submit it (make sure the problem is labeled).

If you have any questions address them to:

- Catia Silva (TA) – catiaspsilva@ufl.edu
- Sheng Zou (TA) – shengzou@ufl.edu

Question 1 - 10 points

Assuming a univariate Gaussian data likelihood given N i.i.d. data points:

$$p(\mathbf{X}|\mu) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) \tag{1}$$

and a Gaussian prior distribution on the mean:

$$p(\mu|\mu_0) = \mathcal{N}(\mu|\mu_0, \sigma_0^2) \tag{2}$$

with fixed variances (σ^2, σ_0^2, and $\sigma^2 \neq \sigma_0^2$), using the method of completing the square (in the exponent) show that the posterior distribution is given by:

$$p(\mu|X) = \mathcal{N}(\mu|\mu_N, \sigma_N^2) \tag{3}$$

where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} \tag{4}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \tag{5}$$

where μ_{ML} is the maximum likelihood solution for μ given the N data points.

Show all your work in the derivation.

Solution for Question 1

$$\begin{aligned}
p(\mu|\mathbf{X}) &\propto p(\mathbf{X}|\mu)p(\mu) \\
&= \prod_{i=1}^{N} \mathcal{N}(x_i|\mu, \sigma^2)\mathcal{N}(\mu|\mu_0, \sigma_0^2) \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right\} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{\sum_{i=1}^{N}\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right) - \frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right)\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2}\left(\frac{\sum_{i=1}^{N} x_i^2 - 2\sum_{i=1}^{N} x_i\mu + \mu^2 N}{\sigma^2} + \frac{\mu^2 - 2\mu\mu_0 + \mu_0^2}{\sigma_0^2}\right)\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi\sigma^2}\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2}\left(\mu^2\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right) - 2\mu\left(\frac{\sum_{i=1}^N x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right) + \frac{\sum_{i=1}^N x_i^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}\right)\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2}\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\mu^2 - 2\mu\left(\frac{\sum_{i=1}^N x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}\right)\right\} \\
&\quad \exp\left\{\frac{\sum_{i=1}^N x_i^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2}\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\mu - \left(\frac{\sum_{i=1}^N x_i\sigma_0^2+\mu_0\sigma^2}{\sigma^2\sigma_0^2}\right)\left(\frac{N\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}\right)^{-1}\right)^2\right. \\
&+ \left.\frac{1}{2}\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\left(\frac{\sum_{i=1}^N x_i\sigma_0^2+\mu_0\sigma^2}{\sigma^2\sigma_0^2}\right)\left(\frac{N\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}\right)^{-1}\right)^2\right\} \exp\left\{\frac{\sum_{i=1}^N x_i^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2}\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\mu - \frac{\sum_{i=1}^N x_i\sigma_0^2+\mu_0\sigma^2}{N\sigma_0^2+\sigma^2}\right)^2\right\} \\
&\quad \exp\left\{\frac{1}{2}\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\frac{\sum_{i=1}^N x_i\sigma_0^2+\mu_0\sigma^2}{N\sigma_0^2+\sigma^2}\right)^2 + \frac{\sum_{i=1}^N x_i^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}\right\} \\
&= C\exp\left\{-\frac{1}{2}\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\mu - \frac{\sum_{i=1}^N x_i\sigma_0^2+\mu_0\sigma^2}{N\sigma_0^2+\sigma^2}\right)^2\right\} \\
&\propto \mathcal{N}\left(\mu \middle| \frac{\sum_{i=1}^N x_i\sigma_0^2+\mu_0\sigma^2}{N\sigma_0^2+\sigma^2}, \left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}\right)
\end{aligned}
$$

Question 2 - 10 points

Extend your solution to Question 1 to a multivariant Gaussian likelihood and Gaussian prior. Assuming a fixed covariance on the prior and the likelihood, derive the MAP solution for the mean vector given N i.i.d. data points and a multivariate Gaussian prior on the mean.

Solution for Question 2

$$
\begin{aligned}
p(\mu|\mathbf{X}) &\propto p(\mathbf{X}|\mu)p(\mu) \\
&= \prod_{i=1}^N \mathcal{N}(\mathbf{x}_i|\mu,\Sigma)\mathcal{N}(\mu|\mu_0,\Sigma_0)
\end{aligned}
$$

$$
\begin{aligned}
&= C_0 \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{N}(\mathbf{x}_i-\mu)^T\Sigma^{-1}(\mathbf{x}_i-\mu)+(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)\right)\right\} \\
&= C_0 \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{N}\mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i-2\sum_{i=1}^{N}\mu^T\Sigma^{-1}\mathbf{x}_i+N\mu^T\Sigma^{-1}\mu\right.\right. \\
&\quad \left.\left.+\mu^T\Sigma_0^{-1}\mu-2\mu^T\Sigma_0^{-1}\mu_0+\mu_0^T\Sigma_0^{-1}\mu_0\right)\right\} \\
&= C_1 \exp\left\{-\frac{1}{2}\left(-2\sum_{i=1}^{N}\mu^T\Sigma^{-1}\mathbf{x}_i+N\mu^T\Sigma^{-1}\mu+\mu^T\Sigma_0^{-1}\mu-2\mu^T\Sigma_0^{-1}\mu_0\right)\right\} \\
&= C_1 \exp\left\{-\frac{1}{2}\mu^T\left(\Sigma_0^{-1}+N\Sigma^{-1}\right)\mu-\frac{1}{2}\mu^T\left(-2\sum_{i=1}^{N}\Sigma^{-1}\mathbf{x}_i-2\Sigma_0^{-1}\mu_0\right)\right\} \\
&= C_1 \exp\left\{-\frac{1}{2}\mu^T\left(\Sigma_0^{-1}+N\Sigma^{-1}\right)\mu+\mu^T\left(\sum_{i=1}^{N}\Sigma^{-1}\mathbf{x}_i+\Sigma_0^{-1}\mu_0\right)\right\} \\
&= C_1 \exp\left\{-\frac{1}{2}\left(\mu^T\left(\Sigma_0^{-1}+N\Sigma^{-1}\right)\mu-2\mu^T\left(\sum_{i=1}^{N}\Sigma^{-1}\mathbf{x}_i+\Sigma_0^{-1}\mu_0\right)\right)\right\} \\
&= C_2 \exp\left\{-\frac{1}{2}\left(\mu-\left(\Sigma_0^{-1}+N\Sigma^{-1}\right)^{-1}\left(\sum_{i=1}^{N}\Sigma^{-1}\mathbf{x}_i+\Sigma_0^{-1}\mu_0\right)\right)^T\right. \\
&\quad \left.\left(\Sigma_0^{-1}+N\Sigma^{-1}\right)^{-1}\left(\mu-\left(\Sigma_0^{-1}+N\Sigma^{-1}\right)^{-1}\left(\sum_{i=1}^{N}\Sigma^{-1}\mathbf{x}_i+\Sigma_0^{-1}\mu_0\right)\right)\right\}
\end{aligned}
$$

Question 3 - 10 points

- In our Binomial/Beta example in class, we computed the ML and MAP solutions for the μ parameter of the Binomial distribution iteratively with an increasing number of trials/random draws. Recall, the parameter μ represented the probability of heads.
- In this homework question, you will do the same sort of experiment for a random draws from a Gaussian distribution (i.e., a Gaussian data likelihood) with a Gaussian prior distribution on the mean parameters (assume a fixed known variance for the Gaussian likelihood and Gaussian prior).
- Using your solution to Question 1, write a script that iteratively draws one data point from the true Gaussian distribution (with known mean). Each iteration compute and ML solution and the MAP solution for the Gaussian mean. After each draw, update the prior distribution to be replaced with the posterior distribution from the previous draw (just like the Binomal/Beta example in class).

- In your solution, provide:
  - Display multiple sample runs of your code and include a description of what the code shows you about ML vs MAP solutions. Your discussion should illustrate that you understand ML and MAP concepts and their differences. Your discussion should answer, at a minimum, the following questions:
    * What happens when the prior mean is initialized to the wrong value? to the correct value?
    * What happens as you vary the prior variance from small to large?
    * What happens when the likelihood variance is varied from small to large?
    * How do the initial values of the prior mean, prior variance, and likelihood variance interact to effect the final estimate of the mean?

Solution for Question 3

Maximum a posteriori (MAP) is an estimate of the underlying model where the prior is incorporated. The prior represents our best guess when lacking of data. In this question, you are asked to code 1) the ML for estimating the Gaussian mean of univariate Gaussian distribution; 2) the MAP for estimating the Gaussian mean of univariate Gaussian distribution, where the prior for Gaussian mean is assumed to be another Gaussian distribution as well; 3) Replace the parameters of prior of Gaussian mean (μ_0. σ_0 is optional) with the MAP estimated parameters of posteriori of Gaussian mean (μ_N. σ_N is optional).

Note that you are asked to draw one data point from the true Gaussian, compute the MAP solution and update the prior distirbution in EACH iteration. Therefore, this is an online learning, i.e. for each iteration when computing Equation (4) and (5), the N is 1 since only one sample is drawn. However, if you would like to use batching learning, the MAP solution is computed one time only using all N samples and prior mean is updated also one time only.

When prior is initialized to the wrong value, you will need more iterations to correct the prior mean to the true mean, compared to the case that prior is initialized close to the true mean. Generally speaking, the prior mean is not easy to be perfectly the true mean. Since the prior variance represents the certainty/uncertainty about the prior mean. When the prior variance is small, we are more certain about the prior mean such that we need more samples (more iterations) to correct the prior mean. Since the likelihood variance is small, the samples we draw from the underlying sample distribution are densely distributed around the true mean, which yields the MAP estimation faster (less iterations) to converge.