CS 676 Machine Learning Homework 1

Department of Applied Mathematics and Statistics

March 3, 2015

1 Bayesian Linear Regression

You work at a prestigious investment bank, and your job as an analyst is to build a model to predict the price of company Y's stock given the current prices for companies A, B, C and D. You will be receiving new price data every millisecond. You would like to build a model that continuously updates its beliefs about ω as you receive new data. You think that a Gaussian distribution approximately captures your beliefs a priori about the parameter ω. Suppose the prior distribution is parameterized with the mean vector μ_0 and covariance matrix Σ_0. The likelihood model is the standard linear regression likelihood:

$$Y_i \sim X_i^T \omega + \mathcal{N}(0, \sigma^2) \tag{1}$$

a. 10 points: After n observations, what will be our prior distribution over w at time $n+1$ (what is the distribution and what are its parameters)?

By formula (1), let us define noise as $\epsilon \sim \mathcal{N}(0, \sigma^2)$, then we have

$$Y_i = X_i^T \omega + \epsilon.$$

From the conditions above, we suppose that the prior distribution of parameter ω satisfies: $\omega \sim \mathcal{N}(\mu_0, \Sigma_0)$, let us define $\omega_1 = \omega; \mu_1 = \mu_0; \Sigma_1 = \Sigma_0$ as the prior of the 1st step, that is to say: $P(\omega_1) = \mathcal{N}(\omega_1 | \mu_1, \Sigma_1)$. Now let us focus on the 1st step: our goal is to compute the prior of the second step (that is to say, the posterior of ω_1 after one observation), since we have $P(Y_1 | X_1, \omega_1, \sigma^2) = \mathcal{N}(Y_1 | X_1^T \omega_1, \sigma^2)$, then by:

$$P(\omega_1 | X_1, Y_1, \sigma^2) \propto P(\omega_1 | \sigma^2) P(Y_1 | X_1, \omega_1, \sigma^2)$$

$$= \mathcal{N}(\omega_1 | \mu_1, \Sigma_1) \mathcal{N}(Y_1 | X_1^T \omega_1, \sigma^2) = \mathcal{N}(\omega_1 | \mu_1, \Sigma_1)$$

in which

$$\begin{aligned} \Sigma_1 &= \sigma^2 (\sigma^2 \Sigma_0^{-1} + X_1^T X_1)^{-1} \\ \Sigma_1^{-1} &= \Sigma_0^{-1} + \frac{1}{\sigma^2} X_1^T X_1 \\ \mu_1 &= \Sigma_1 \Sigma_0^{-1} \mu_0 + \frac{1}{\sigma^2} \Sigma_1 X_1^T Y_1 \end{aligned}$$

Then let $\omega_2 = \omega_1 | X_1, Y_1, \sigma^2$ with the prior of μ_2 is μ_1 and prior of Σ_2 is Σ_1, therefore, we have:

$$\omega_2 \sim \mathcal{N}(\mu_2, \Sigma_2), \text{ and } P(\omega_2) = \mathcal{N}(\omega_2 | \mu_2, \Sigma_2).$$

Repeat the ways above, we could get the prior of $i+1$th step under the ith step. Now, for the $n+1$th step, the prior of the $n+1$th step is the posterior of nth step, if we have the posterior of nth step as $\omega_n \sim \mathcal{N}(\mu_n, \Sigma_n)$, then by the ways above, we get $\omega_{n+1} \sim \mathcal{N}(\mu_{n+1}, \Sigma_{n+1})$ in which

$$\begin{aligned}
\Sigma_{n+1} &= \sigma^2(\sigma^2\Sigma_n^{-1} + X_n^T X_n)^{-1} \\
\Sigma_{n+1}^{-1} &= \Sigma_n^{-1} + \frac{1}{\sigma^2} X_n^T X_n \\
\mu_{n+1} &= \Sigma_{n+1}\Sigma_n^{-1}\mu_n + \frac{1}{\sigma^2}\Sigma_{n+1} X_n^T Y_n
\end{aligned}$$

b. 5 points: Given X_{n+1} and the prior density you computed above, write down the full posterior predictive distribution over Y_{n+1}.

By formula (7.60) in Murphy's book, the posterior predictive distribution over Y_{n+1} could be shown as:

$$P(Y_{n+1}|X_{n+1}, \omega_{n+1}, \sigma^2) = \int \mathcal{N}(Y_{n+1}|X_{n+1}^T \omega_{n+1}, \sigma^2)\mathcal{N}(\omega_{n+1}|\mu_{n+1}, \Sigma_{n+1}) d\omega_{n+1}$$

$$= \mathcal{N}(Y_{n+1}|\omega_{n+1}^T X_{n+1}, \sigma_{n+1}^2(X_{n+1}))$$

where

$$\sigma_{n+1}^2(X_{n+1})) = \sigma^2 + X_{n+1}^T \Sigma_{n+1,N} X_{n+1}$$

c. 5 points: Suppose we are unsure about our choice of hyperparameters μ_0 and σ_0. To account for our uncertainty, we can use hierarchical Bayes and place some prior $\pi(\mu_0, \Sigma_0)$ over the hyperparameters. What estimation technique could we use to approximate the prior in the full hierarchical Bayes model?

To account for our uncertainty of hyperparameter, for the first step, and the posterior of the first step could be

$$P(\pi, \omega_1|X_1, Y_1, \omega_1) \propto P(Y_1|X_1, \omega_1, \sigma)P(\omega_1|\mu_1, \Sigma_1)\pi(\mu_0, \Sigma_0).$$

Then for each step(say step i), we **do not** directed define the prior of ith step ω_i as the posterior of $i-1$th step (say $(\omega_{i-1}|X_{i-1}, Y_{i-1}, \sigma)$); Instead, we define the prior of ith step as this:
(i) We have distribution of the posterior of $i-1$th step as

$$(\omega_{i-1}|X_{i-1}, Y_{i-1}, \sigma) = \omega_{i-1,\text{posterior}} \sim \mathcal{N}(\mu_{i-1,\text{posterior}}, \Sigma_{i-1,\text{posterior}})$$

(ii) Then define the distribution of ith step as

$$\omega_i \sim \mathcal{N}(\mu_i, \Sigma_i)$$

in which μ_i, Σ_i are random variables, which satisfies:

$$(\mu_i, \Sigma_i) \sim \pi(\mu_{i-1,\text{posterior}}, \Sigma_{i-1,\text{posterior}})$$

Therefore, we have the probability of ith prior $P(\omega_i)$ could be computed as

$$P(\omega_i) = \int P(\omega_i|\mu_i, \Sigma_i)\pi(\mu_i, \Sigma_i|\mu_{i-1,\text{posterior}}, \Sigma_{i-1,\text{posterior}}) d\mu_i d\Sigma_i$$

Then the posterior of ith step $\omega_i|X_i, Y_i, \sigma$ is

$$P(\pi, \omega_i|X_i, Y_i, \sigma) \propto P(Y_i|X_i, \omega_i, \sigma)P(\omega_i)$$

$$\propto P(Y_i|X_i, \omega_i, \sigma)P(\omega_i|\mu_i, \Sigma_i)\pi(\mu_i, \Sigma_i|\mu_{i-1,\text{posterior}}, \Sigma_{i-1,\text{posterior}})$$

In approximating the prior in the full hierarchical Bayes model, by Empirical Bayes or evidence approximate, we could find the hyperparameters maximizes the marginal likelihood as

$$\hat{\pi} = \arg\max\{\int P(Y|X, \omega, \sigma)P(\omega|\pi) d\omega\}$$

d. 5 points: You need to be able to predict Y_{n+1} quickly, why is the full posterior predictive distribution a bad choice in this scenario? How could you approximate it? Under what condition is this approximation reasonable?

Since full posterior predictive distribution requires to marginalize all the parameters, which is hard to compute. Since by part(c) above, we could compute $\hat{\pi}$ as an approximation, then we have $\hat{\pi} = \hat{\pi}(\hat{\mu_0}, \hat{\Sigma_0})$, then for the posterior of ω we have

$$P(\omega|Y) = P(Y|\omega)P(\omega|\hat{\mu_0}, \hat{\Sigma_0})$$

And the condition of this approximation reasonable is that if the posterior is sharply peaked around our $\hat{\pi}$.

e. 10 points: After observing 10,000 examples, your boss comes to let you know that the PhDs in the back have suggested that there may be a correlation between the influences that the stock prices for companies A and B or for companies C and D have on Y's price, but they're not sure which. In fact, they're not even entirely sure that the correlation exists.
Let $\mathcal{M}_0$ denote the model in which there are no correlations between any of the weights in the model, $\mathcal{M}_{AB}$ denote the model in which the weights for A and B are correlated, and $\mathcal{M}_{CD}$ denote the model in which the weights for C and D are correlated. Assuming that all variances are σ_0^2 and all covariances are γ_0^2, how would you encode the beliefs of these three models in three different prior distributions?

These three models can show the correlation between the free weight parameter in ω. $\mathcal{M}_0$ means the four parameters are independent on each other. $\mathcal{M}_AB$ means A,B has somewhat correlation, and $\mathcal{M}_CD$ means C,D has somewhat correlation. Therefore, there exist three forms of initial Σ_0:

$$\Sigma_{0,AB} = \begin{pmatrix} \sigma_0^2 & \gamma_0^2 & 0 & 0 \\ \gamma_0^2 & \sigma_0^2 & 0 & 0 \\ 0 & 0 & \sigma_0^2 & 0 \\ 0 & 0 & 0 & \sigma_0^2 \end{pmatrix}$$

$$\Sigma_{0,CD} = \begin{pmatrix} \sigma_0^2 & 0 & 0 & 0 \\ 0 & \sigma_0^2 & 0 & 0 \\ 0 & 0 & \sigma_0^2 & \gamma_0^2 \\ 0 & 0 & \gamma_0^2 & \sigma_0^2 \end{pmatrix}$$

$$\Sigma_{0,\text{ind}} = \begin{pmatrix} \sigma_0^2 & 0 & 0 & 0 \\ 0 & \sigma_0^2 & 0 & 0 \\ 0 & 0 & \sigma_0^2 & 0 \\ 0 & 0 & 0 & \sigma_0^2 \end{pmatrix}$$

f. 20 points: You'd like to compare the models by computing the posterior distribution over $\mathcal{M}_0$, $\mathcal{M}_{AB}$ and $\mathcal{M}_{CD}$. Recall that the posterior distribution over model $\mathcal{M}_i$ is

$$P(\mathcal{M}_i|y_{1:n}, x_{1:n}) \propto P(y_{1:n}|x_{1:n}, \mathcal{M}_i)P(\mathcal{M}_i) \tag{2}$$

Derive a closed form expression for $P(y_{1:n}|x_{1:n}, \mathcal{M}_i)$ where Σ_i is the covariance matrix associated with model $\mathcal{M}_i$.

By formula (2), and the parameter given from part (e), we have

$$P(\mathcal{M}_i|y_{1:n}, x_{1:n}) \propto \prod_{i=1}^{n} \mathcal{N}(y_i|x_i^T\omega_i, \sigma_0^2)\mathcal{N}(\omega_i|\mu_i, \Sigma_i)$$

	μ	Σ
M_0	$\begin{pmatrix} 0.57070445 \\ 0.25711002 \\ 0.21197589 \\ 0.71122906 \end{pmatrix}$	$\begin{pmatrix} 7.89425278e-07 & -2.80370998e-09 & 3.31753511e-07 & -2.88579106e-07 \\ -2.80370998e-09 & 5.19173768e-07 & 2.11978748e-07 & 1.63336624e-07 \\ 3.31753511e-07 & 2.11978748e-07 & 2.85664750e-07 & -1.11808635e-07 \\ -2.88579106e-07 & 1.63336624e-07 & -1.11808635e-07 & 3.28703746e-07 \end{pmatrix}$
M_{AB}	$\begin{pmatrix} 0.57070444 \\ 0.25711017 \\ 0.21197595 \\ 0.71122911 \end{pmatrix}$	$\begin{pmatrix} 7.89425067e-07 & -2.80343552e-09 & 3.31753535e-07 & -2.88578943e-07 \\ -2.80343552e-09 & 5.19173676e-07 & 2.11978826e-07 & 1.63336495e-07 \\ 3.31753535e-07 & 2.11978826e-07 & 2.85664792e-07 & -1.11808619e-07 \\ -2.88578943e-07 & 1.63336495e-07 & -1.11808619e-07 & 3.28703646e-07 \end{pmatrix}$
M_{CD}	$\begin{pmatrix} 0.57070461 \\ 0.25711009 \\ 0.21197602 \\ 0.71122898 \end{pmatrix}$	$\begin{pmatrix} 7.89425085e-07 & -2.80372237e-09 & 3.31753389e-07 & -2.88578968e-07 \\ -2.80372237e-09 & 5.19173790e-07 & 2.11978749e-07 & 1.63336648e-07 \\ 3.31753389e-07 & 2.11978749e-07 & 2.85664676e-07 & -1.11808541e-07 \\ -2.88578968e-07 & 1.63336648e-07 & -1.11808541e-07 & 3.28703656e-07 \end{pmatrix}$

Table 1: posterior distribution over the three possible models on 10000th step

$$\propto \prod_{i=1}^{n} P(\omega_i|x_i, y_i) = \propto \prod_{i=1}^{n} \mathcal{N}(\omega_{i+1}|\mu_{i+1}, \Sigma_{i+1})$$

The formula above is gotten from part (a), in which

$$\Sigma_{i+1} = \sigma_0^2(\sigma_0^2\Sigma_i^{-1} + x_i x_i^T)^{-1}$$
$$\Sigma_{i+1}^{-1} = \Sigma_i^{-1} + \frac{1}{\sigma_0^2} x_i x_i^T$$
$$\mu_{i+1} = \Sigma_{i+1}\Sigma_i^{-1}\mu_i + \frac{1}{\sigma_0^2}\Sigma_{i+1} x_i y_i$$

g. 20 points: Using the data distributed with homework 1 (stocks.csv available on Piazza), compute the posterior distribution over the 3 models. Assume a uniform prior over the models, and let $\mu_0 = 0, \sigma^2 = 4, \sigma_0^2 = 1$ and $\gamma_0^2 = 1/2$. Include a table showing the posterior distribution over the three possible models. Which one would you choose?

We implement computation by Python, with $\mu_0 = 0, \sigma^2 = 4, \sigma_0^2 = 1, \gamma_0^2 = 1/2$, and *stocks.csv*. Table 1 show the posterior distribution over the three possible models on the 10000 step.

Figure 1 shows the prediction of y by using M_0, M_{AB}, M_{CD}. We also compute the loss function by using mean squared error MSE of each model, where

$$MSE = \sum_{t=1}^{10000} (y_{t,real} - X_t^T \mu_t)^2$$

we obtained the average of MSE by

$$avgMSE = \frac{MSE}{10000}$$

The result shows

$$avgMSE(M_0) = 72.57644754, avgMSE(M_{AB}) = 72.5764736, avgMSE(M_{CD}) = 72.57657563$$

We can find that

$$avgMSE(M_0) \approx avgMSE(M_{AB}) \approx avgMSE(M_{CD})$$

Therefore, M_0, M_{AB}, M_{CD} are approximately equivalent to each other.

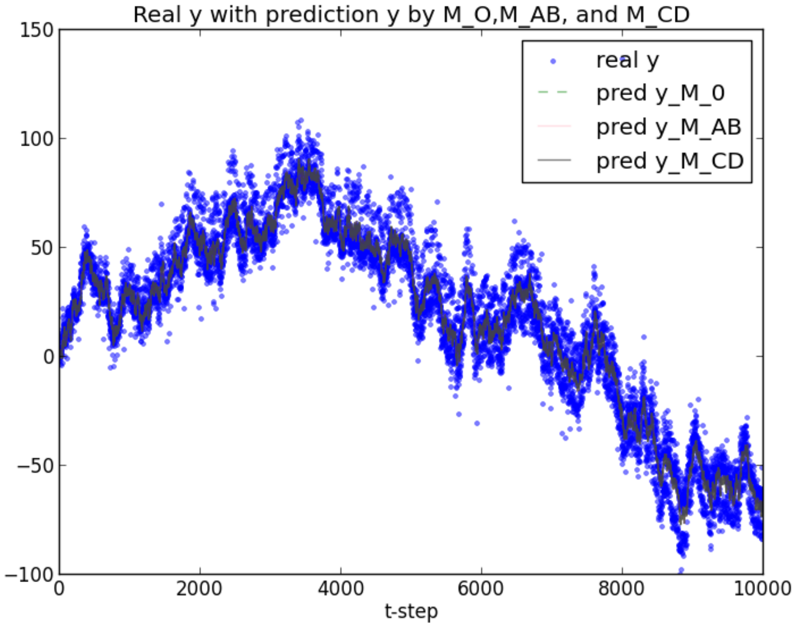

Figure 1: Real y with prediction y by M_0,M_{AB},M_{CD}

h. 15 points: Recent issues with the company's communications infrastructure has caused some unusually noisy observations to be collected. Assuming that the error for observation i is now defined by

$$\epsilon_i \sim \theta\mathcal{N}(0,\sigma^2) + (1-\theta)\mathcal{N}(0,50) \tag{3}$$

Where $0 < \theta < 1$, what is the new posterior distribution over ω under this noise model at time n?

Since ϵ_i is shown as formula (3), then by the property of linear combination of Gaussian distribution we have

$$\epsilon_i \sim \mathcal{N}(\mu^*, (\sigma^*)^2)$$

in which

$$\mu^* = 0$$
$$(\sigma^*)^2 = \theta^2\sigma^2 + (1-\theta)^2 50$$

Then the new posterior distribution over ω under this noise model at time n is

$$\mathcal{N}(Y_n | X_n^T\omega_n, (\sigma^*)^2 + X_n^T\Sigma_n X_n)$$