Quiz 1

14 Marks, 15% weightage, 45 minutes

Saturday 21st January, 2017

Please answer **to the point** in the limited space provided for each question. You can do rough work in a separate sheet of paper provided to you. You can also assume any result stated or proved in the class (but NOT as part of the tutorials).

Problem 1. A student of machine learning measured the height of each student in a school along with his/her age, weight and the height of the child's mother and father. The student then fitted a linear regression model to predict the height as a function of the the other observations:

height = f(age, weight, height of the child's mother, height of the child's father)

The student of Machine Learning suddenly realizes that she had somehow measured the height of each child with his/her shoe on. This means the learnt model had to be corrected now. It was known that every child in the school wore a shoe of 1.5 mm thickness. What is the simplest correction the ML student can do to the model (without computing it all over again) to get the same result as she would have obtained by using the correct height (excluding the shoe thickness)?

Prove your answer.

(2.5 Marks)

Solution:

Consider

1. $\hat{\mathbf{w}_{ML}} = \text{argmin} \sum_{j=1}^{m} (\mathbf{w}^T \phi(\mathbf{x}_j) + b - y_j)^2$

2. $\hat{\mathbf{w}_{ML}} = \text{argmin} \sum_{j=1}^{m} (\mathbf{w}^T \phi(\mathbf{x}_j) + b_{new} - (y_j + 1.5))^2$

Claim is $b_{new} = b + 1.5$, where one can obtain the solution b_{new} to (2) simply by subtracting 1.5 mm from b. If not, then one can show that the solution $b_{new} \neq b + 1.5$ to the second should have yielded a better solution $b_{new} - 1.5 \neq b$ to (1) - which is a contradiction.

Problem 2. Consider real-valued variable $X \in \Re$. A random variable Y is generated, conditioned on X, based on the following process:

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$Y = aX + \epsilon$$

Assume we have a training dataset of m pairs (x_i, y_i) for $i = 1..m$, and σ is known. Analytically derive the correct expression for the maximum likelihood estimate of a.

(5.5 Marks)

Solution:

This is a very special case of maximum likelihood estimation for linear regression with $\phi(\mathbf{x}) \in \Re$ (that is, with $p = 1$). This can be also be solved by using $\mathbf{w} = (\mathbf{\Phi^T \Phi})^{-1} \mathbf{\Phi^T y}$ and substituting $\Phi = [x_1; x_2; ...x_m]$ and $\mathbf{w} = [a]$ to get $a = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$

Another way of proving $a = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$ is from first principles. Here is the second method:

Solve for $\arg\max_a \prod_i exp\big(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\big)$. Equivalently, you could also solve for maximizing the monotonically increasing (log) transformation of the objective

$$\arg\max_a \log\left(\prod_i exp\big(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\big)\right) = \sum_i \left(\big(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\big)\right)$$

Taking partial derivative w.r.t. a we get $\sum_i ax_i^2 - x_i y_i = 0$. That is, $a = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$

Problem 3. In class, we have illustrated Bayesian estimation for the parameter μ of a Normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, assuming that σ was known by imposing a Normal (conjugate) prior on μ. Now suppose that the parameter μ is known and we wish to estimate σ^2. What will be the form of the conjugate prior for this estimation procedure? If $\mathcal{D} = X_1, X_2, X_3, ..., X_n$ is a set of independent samples from this distribution, after imposing the conjugate prior, compute the form of the likelihood function $\mathcal{L}(\theta)$, the posterior density $P(\theta \mid \mathcal{D})$ and the posterior probability $P(X \mid \mathcal{D})$. Again, you can ignore normalization factors

(6 Marks)
Solution:

- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 ... x_m$
- $\mu_{MLE} = \frac{1}{m}\sum_{i=1}^{m} x_i$ and $\sigma^2_{MLE} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{MLE})^2$
- Suppose you are told that σ^2 is a random variable is and μ is not.

$$\Pr(x_i|\mu;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\Pr(\mathcal{D}|\mu) = \left(\frac{1}{(2\pi)^{\frac{m}{2}}(\sigma^2)^m}\right) exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2\right)$$

Note the positions of σ^2 in the likelihood above. In order to make the posterior of the same form as the prior, we should make the prior jell seamlessly with the likelihood because $\Pr(\theta|\mathcal{D}) \propto \Pr(\mathcal{D}|\theta)\Pr(\theta)$, where we have $\theta := \sigma^2$. This could mean

$$\Pr(\theta) \propto \frac{1}{\theta^A} exp(\frac{-B}{\theta})$$

One can normalize this distribution to find the proportionality constant.

(TAs. It is ok if the students get so far in suggesting the prior. It is also ok if the students miss somehow land up only with $\Pr(\theta) \propto \frac{1}{\theta} exp(\frac{-B}{\theta})$)

Part 2 of the question:

$$\Pr(x|D) = \int_{\theta} \Pr(x|\theta)\Pr(\theta|D)d\theta$$

Substituting,

$$\Pr(x|D) = \int_{\sigma^2} \Pr(x|\sigma^2)\Pr(\sigma^2|D)d\sigma^2 = \int_{\sigma^2} \Pr(x|\sigma^2)\Pr(\sigma^2|D)d\sigma^2$$

We can substitute and leave the integral as it is. An approximation is to use the MAP or Bayes estimate in place of integration and

$\Pr(x|D) \approx \Pr(x|\sigma^2_{\mathbf{MAP}})\Pr(\sigma^2_{\mathbf{MAP}}|D)$

No need to give marks to what follows: This is called an inverse-gamma distribution.

$$p(\theta) = \frac{B^{A-1}}{\Gamma(A-1)}\frac{1}{\theta^A} exp(\frac{-B}{\theta})$$

The posterior is also an inverse-gamma distribution with

$$A' = A + n/2$$

$$B' = B + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Marginal likelihood

$$p(\mathcal{D}) = \frac{\frac{1}{\sqrt[2]{2\pi\theta}} exp(\sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2\theta}) \frac{B^{A-1}}{\Gamma(A-1)} \frac{1}{\theta^A} exp(\frac{-B}{\theta})}{\frac{B'^{A'-1}}{\Gamma(A'-1)} \frac{1}{\theta^{A'}} exp(\frac{-B'}{\theta})}$$

Posterior Predictive

$$p(x|\mathcal{D}) = \frac{p(x, \mathcal{D})}{p(\mathcal{D})}$$

Use $\mathcal{D}' = (\mathcal{D}, x)$ to compute $p(\mathcal{D}')$ and substitute back to get

$$p(x|\mathcal{D}) = t_{2A'}(x|\mu, \theta = \frac{B'}{A'})$$