Data Science 423
Homework 5

Dennis Lee

November 26, 2019

1. For Bayesian Linear Regression, we make use of Bayes' Theorem:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Normalization}}$$

*I'm using the books notation.

Where our data is X & $X^T\varphi$ & $\Theta = [\vec{\varphi}, \sigma^2]$

$$Pr(\varphi \mid X, \vec{w}) = \frac{Pr(\vec{w} \mid X, \vec{\varphi})\ P(\varphi)}{Pr(\vec{w} \mid X)}$$

We model the prior as a multivariate normal distribution:

$$\text{Prior: } Pr(\varphi) = N_{\vec{\varphi}}[0, \sigma_p^2 I]$$

For the likelihood with we write the likelihood as:

$$\text{Likelihood: } Pr(\vec{w} \mid X, \Theta) = N_{\vec{w}}[X^T\vec{\varphi}, \sigma^2 I]$$

Focusing on the numerator for now:

$$Pr(\varphi \mid X, \vec{w}) \propto N_{\vec{w}}[X^T\vec{\varphi}, \sigma^2 I]\ N_{\vec{\varphi}}[0, \sigma_p^2 I]$$

we then make use of several relations.

$$Pr(\varphi|X,\vec{\omega}) \propto \underbrace{N_{\vec{\omega}}[X^T\vec{\varphi}, \sigma^2 I]}\, N_{\vec{\varphi}}[\emptyset, \sigma_p^2 I]$$

Using: $N_x[Ay+b, \Sigma] = \kappa \cdot N_y[A'x+b', \Sigma']$

we find that in our case:

$$\Sigma = \sigma^2 I \longrightarrow \Sigma' = (X\sigma^{-2}X^T)^{-1}$$

$$A = X^T \longrightarrow A' = (X\sigma^{-2}X^T)^{-1} X\sigma^{-2}$$

$$b = \emptyset \longrightarrow b' = \emptyset$$

Thus:

$$Pr(\varphi|X,\vec{\omega}) \propto N_{\vec{\varphi}}[(\sigma^{-2}XX^T)^{-1}X\sigma^{-2}\omega, (\sigma^{-2}XX^T)^{-1}] \times N_{\vec{\varphi}}[\emptyset, \sigma_p^2 I]$$

Combining using:

$$N_x[a,A]N_x[b,B] = \kappa\, N_x[(A^{-1}+B^{-1})^{-1}(A^{-1}a+B^{-1}b), (A^{-1}+B^{-1})^{-1}]$$

For us: $a = (\sigma^{-2}XX^T)^{-1}X\sigma^{-2}\omega \quad b = \emptyset$

$A = (\sigma^{-2}XX^T)^{-1} \quad B = \sigma_p^2 I$

$$Pr(\varphi|X,\omega) \propto \kappa\, N_\varphi\Big[\big[(\sigma^{-2}XX^T) + \sigma_p^2 I\big]^{-1}\big((\sigma^{-2}XX^T)(\sigma^{-2}XX^T)^{-1}X\sigma^{-2}\omega\big), \big[(\sigma^{-2}XX^T) + \sigma_p^2 I\big]^{-1}\Big]$$

$$Pr(\varphi|X,\omega) \propto \kappa N_\varphi\left[\left[(\sigma^{-2}XX^T)+(\sigma_p^2 I)\right]^{-1}\left((\sigma^{-2}XX^T)\ (\sigma^{-2}XX^T)^{-1}X\sigma^{-2}\omega\right), \left[(\sigma^{-2}XX^T)+(\sigma_p^2 I)\right]^{-1}\right]$$

$$\propto \kappa N_\varphi\left[\underbrace{\left[(\sigma^{-2}XX^T)+\sigma_p^{-2}I\right]}^{-1}\left(X\sigma^{-2}\omega\right), \underbrace{\left[(\sigma^{-2}XX^T)+\sigma_p^{-2}I\right]}^{-1}\right]$$

If we define $A = \sigma^{-2}XX^T + \sigma_p^{-2}I$

$$Pr(\varphi|X,\omega) \propto N_\varphi\left[\sigma^{-2}A^{-1}X\omega,\ A^{-1}\right]$$

$$Pr(\varphi|X,\omega) \propto N_\varphi\left[\sigma^{-2}A^{-1}X\omega, A^{-1}\right] \text{ where } A = \sigma^{-2}XX^T + \sigma_p^{-2}I$$

Since I used the book notation, to use the notation of the homework, we get:

$$Pr(\tilde{\omega}|\tilde{X},\bar{y}) = N_{\tilde{\omega}}\left[\sigma^{-2}A^{-1}\tilde{X}\bar{y}, A^{-1}\right]$$

$$A = \sigma^{-2}\tilde{X}\tilde{X}^T + \sigma_p^{-2}I$$

2

To show that:

$$Pr\left(y^* \mid \bar{x}^*, \tilde{X}, \bar{y}\right) = \mathcal{N}_{y^*}\left[\sigma^{-2}\bar{x}^{*T}A^{-1}\tilde{X}\bar{y},\ \bar{x}^{*T}A^{-1}\bar{x}^* + \sigma^2\right]$$

we start with the integral:

$$Pr(y^* \mid \bar{x}^*, \tilde{X}, \bar{y}) = \int \underbrace{\mathcal{N}_{y^*}\left[\tilde{w}^T x^*, \sigma^2\right]}_{\text{predictions of each } \tilde{w}} \underbrace{\mathcal{N}_{\tilde{w}}\left[\frac{1}{\sigma^2}A^{-1}\tilde{X}\bar{y}, A^{-1}\right]}_{\text{posterior distribution}} d\tilde{w}$$

Since each of the predictions are Gaussian, we can argue that the predictive distribution will also likewise be gaussian. Alternatively, we can see that if we go through all the steps, as in problem 1, we will recieve a constant defined as a Gaussian when computing the product of two normals. Either way, knowing this, we simply compute the mean (μ_{pred}) & variance (σ^2_{pred}) to define the Gaussian

Looking at the integral, we see that the expectation value/mean of y^* is $\bar{x}^*$ multiplied by the expectation value of the weight $\tilde{w}$. Since that's just the mean:

$$\mu_{pred} = E[y^*] = E[\tilde{w}^T x^*] = x^{*T} E[\tilde{w}]$$

$$= \bar{x}^{*T}\left(\frac{1}{\sigma^2}A^{-1}\tilde{X}\bar{y}\right)$$

From the Computer Vision textbook (Ch4) we can write:

$$\sigma^2_{pred} = \sigma^2 + \bar{x}^{*T} A^{-1} \bar{x}^*$$

Combining all of this, we have:

$$p(y^* | \bar{x}^*, \tilde{X}, \bar{y}) = N_{y^*}\left[\sigma^{-2} \bar{x}^{*T} A^{-1} \tilde{X} \bar{y},\ \bar{x}^{*T} A^{-1} \bar{x}^* + \sigma^2\right]$$

3// Given the logistic regression model:

$$L = \sum_{i=1}^{P} y_i \log\left[\sigma(a_i)\right] + \sum_{i=1}^{P} (1-y_i)\log\left[1-\sigma(a_i)\right]$$

where $a_i = \bar{x}_i^T \tilde{\omega}$, the gradient with respect to $\tilde{\omega}$ is given by:

$$\nabla_{\tilde{\omega}} L = \sum_{i=1}^{P} y_i \frac{1}{\sigma(a_i)} \nabla_w \sigma(a_i) + \sum_{i=1}^{P} (1-y_i) \frac{1}{1-\sigma(a_i)} \left[-\nabla_w \sigma(a_i)\right]$$

$$= \sum_{i=1}^{P} y_i (1-\sigma(a_i)) \nabla_w a_i + \sum_{i=1}^{P} (1-y_i)(-1)\sigma(a_i) \nabla_w a_i$$

$$= \sum_{i=1}^{P} y_i (1-\sigma(a_i)) \bar{x}_i + \sum_{i=1}^{P} (1-y_i)(-1)\sigma(a_i) \bar{x}_i$$

$$= \sum_{i=1}^{P} \left[-1(-y_i + \sigma(a_i) y_i) + (-1)(\sigma(a_i) - \sigma(a_i) y_i)\right] \bar{x}_i$$

$$= -\sum_{i=1}^{P} \left(-y_i + \sigma(a_i) y_i + \sigma(a_i) - \sigma(a_i) y_i\right) \bar{x}_i$$

$$= -\sum_{i=1}^{P} \left(-y_i + \sigma(a_i)\right) \bar{x}_i$$

$$\boxed{\nabla_{\tilde{\omega}} L = -\sum_{i=1}^{P} \left(\sigma(a_i) - y_i\right) \bar{x}_i}$$

4

Starting with:

$$\nabla_{\tilde{w}} L = -\sum_{i=1}^{P} (\sigma(a_i) - y_i)\bar{x}_i$$

we find the Hessian:

$$\nabla^2_{\tilde{w}} L = -\sum_{i=1}^{P} \nabla_{\tilde{w}} \sigma(a_i)\bar{x}_i$$

$$= -\sum_{i=1}^{P} \sigma(a_i)(1-\sigma(a_i))\bar{x}_i \nabla_{\tilde{w}} a_i$$

$$= -\sum_{i=1}^{P} \sigma(a_i)(1-\sigma(a_i))\bar{x}_i \bar{x}_i^T$$

$$\boxed{\nabla^2_{\tilde{w}} L = -\sum_{i=1}^{P} \sigma(a_i)\left(1-\sigma(a_i)\right)\bar{x}_i \bar{x}_i^T}$$