IE 7374 ST: Machine Learning in Engineering
HW-2

Rohit Bokade
NUID: 001280767

March 10, 2021

1. (a) The best decision function can be given by $f^*(x) = x$. Thus, the expected loss or "risk" would be 0 $R(f) = \mathbb{E}\ell(f^*(x), y) = 0$.

   (b) Approximation function: $R(f_F) - R(f^*)$. The best constant function would be $f(x) = \mathbb{E}[X] = 5.5$. Thus the approximation error would be $\mathbb{E}[(f^*(X) - 5.5)^2] = \mathbb{E}[(Y - 5.5)^2] = Var(Y) = \frac{33}{4} = 8.25$.

   (c) i. Hypothesis space F of affine functions $f(x) = a + bx$. The best estimation function within this hypothesis space would have the risk of 0. And so, the approximation error would also be 0. Reference

   ii.

$$\hat{f}(x) = x + 1$$
$$R(\hat{f}) - R(f_F) = \mathbb{E}[(Y - X + 1)^2] - 0 = \mathbb{E}[1] = 1$$

2. (a) Since, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, $y \sim \mathcal{N}(w^T x + b, \sigma^2)$. Thus, we can write

$$p(y = y_i | x_i, w_i, b) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\left(\frac{y_i - w_i x_i - b}{\sigma}\right)^2\right)$$

   Reference

   (b)

$$P(y|\beta) = \prod_{i=1}^{N} p(y_i | x_i, w_i, b)$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^N \exp -\frac{1}{2}\left(\frac{\sum_{i=1}^{n}(y_i - w_i x_i - b)^2}{\sigma^2}\right)$$

Taking log on both sides

$$\ln P(y|\beta) = -n \log \sqrt{2\pi\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w_i x_i - b)^2$$

   (c)

$$lnP(y|\beta) = -n \log \sqrt{2\pi\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w_i x_i - b)^2$$

Maximizing the log likelihood is equivalent to minimizing the "summation" in the second term

$$\begin{aligned}\max_{\beta} \ln P(y|\beta) &= \min_{\beta}(\sum_{i=1}^{n}(y_i - w_i x_i - b)^2) \\ &= arg\min_{\beta}(y - x^T\beta)^T(y - x^T\beta)\end{aligned}$$

Other terms are constant.

(d) To derive the values of coefficients β, we can take derivative of the log likelihood with respect to β and equating it to zero.

$$J(\beta) = (y - X\beta)^T((y - X\beta))$$

Taking derivative with respect to β

$$\begin{aligned}\frac{\partial}{\partial\beta}J(\beta) &= 2X^T(X\beta - y) = 0 \\ \therefore X^T(X\beta - y) &= 0 \\ \therefore X^TX\beta - X^Ty &= 0 \\ \therefore \beta &= (X^TX)^{-1}X^Ty\end{aligned}$$

3. (a)

$$\begin{aligned}\hat{\beta} &= arg\min_{\beta}\{(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta\} \\ \frac{\partial\hat{\beta}}{\partial\beta} &= \frac{\partial(Y - X\beta)^T(Y - X\beta)}{\partial\beta} + \frac{\partial\lambda\beta^T\beta}{\partial\beta} \\ &= \frac{Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta}{\partial\beta} + \frac{\lambda\beta^T\beta}{\partial\beta} \\ &= -2X^T(Y - \beta^TX) + 2\lambda\beta \\ \therefore \beta &= (X^TX + \lambda\mathbb{I})^{-2}X^TY \\ &= (A^TA + \lambda\mathbb{I})^{-2}A^TY\end{aligned}$$

Reference

(b) First of all, A^TA is a symmetric matrix. Let A^TA have dimensions $n \times n$. Therefore, for it to be full rank, the rank of the matrix should be n. Further, A^TA will always be a semi-definite matrix and all of its **eigen values would be greater than zero** $v \geq 0 \ \forall \ v \in V$.

Next, any eigen vector $\mathbf{v}_i$ of A^TA will be the eigen vector of $(A^TA + \lambda\mathbb{I})$ scaled as $v_i + \lambda$.

$$\begin{aligned}(A^TA + \lambda\mathbb{I})\mathbf{v}_i &= \underbrace{A^TA\mathbf{v}_i}_{A\mathbf{v}=u\mathbf{v}} + \lambda\mathbb{I}\mathbf{v}_i \\ &= (v_i + \lambda)\mathbf{v}_i\end{aligned}$$

Thus, we can show that the eigen values of $(A^TA + \lambda\mathbb{I}) \geq 0$ and therefore it is full rank and invertible. Reference

4. (a) We can write μ in terms of the training data $x = (x_1, x_2, \cdots, x_n)$ using Bayes' rule

$$P(\mu|x) = \frac{\overbrace{P(x|\mu)}^{\mathcal{N}(\mu,\sigma)} \overbrace{P(\mu)}^{\mathcal{N}(\mu_0,\sigma_0)}}{\underbrace{P(x)}_{constant}}$$

$$P(x|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$P(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

Next, taking log on both sides, we get

$$\log(P(\mu|x)) = \left(\sum_{i=1}^{n} -\log\left(\sqrt{2\pi\sigma^2}\right)\right) - \log\left(\sqrt{2\pi\sigma_0^2}\right) - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}$$

We can take derivative with respect to μ and equate it to 0.

$$\frac{\partial \log(P(\mu|x))}{\partial \mu} = \left(\sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2}\right) - \frac{\mu-\mu_0}{\sigma_0^2} = 0$$

$$\therefore \mu = \frac{\sigma^2\mu_0 + \sigma_0^2 \sum_{i=1}^{n} x_i}{\sigma^2 + n\sigma_0^2}$$

(b) Let $\bar{x}$ be sampled from Gaussian distribution $x_i \sim \mathcal{N}(\mu, \sigma)$. The likelihood function can be written as

$$p(\bar{x_i}|\mu,\sigma) = \prod_{i=1}^{n} f(x_i;\mu,\sigma)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

Next, taking log on both sides, we get

$$p(\bar{x_i}|\mu,\sigma) = \sum_{i=1}^{n} \left(-\log\sqrt{2\pi\sigma^2} - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= \sum_{i=1}^{n} \left(-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{n}{2}\log(2\pi\sigma^2) + \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(x_i-\mu)^2$$

To obtain the estimate for μ we can derivate with respect to μ and equate it to zero.

$$
\begin{aligned}
&\frac{\partial}{\partial \mu}\left(\frac{n}{2}\log(2\pi\sigma^2)+\sum_{i=1}^{n}-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)\\
&=\frac{\partial}{\partial \mu}\left(-\frac{n}{2}\log(2\pi\sigma^2)\right)-\frac{\partial}{\partial \mu}\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\\
&=\frac{\partial}{\partial \mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\\
&=-\frac{1}{2\sigma^2}\frac{\partial}{\partial \mu}\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)\\
&=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2=0\\
\therefore \mu &= \frac{1}{n}\sum_{i=1}^{n}x_i
\end{aligned}
$$

Next, to obtain the estimate for σ^2, we can follow similar procedure.

$$
\begin{aligned}
&\frac{\partial}{\partial \sigma^2}\left(\frac{n}{2}\log(2\pi\sigma^2)+\sum_{i=1}^{n}-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)\\
&=\frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\log(2\pi\sigma^2)\right)-\frac{\partial}{\partial \sigma^2}\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\\
&=-\frac{n}{2}\frac{\partial}{\partial \sigma^2}(\log(2\pi\sigma^2))+\frac{\partial}{\partial \sigma^2}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\\
&=-\frac{n}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2\\
&=\frac{1}{2\sigma^2}\left(-n+\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\\
\therefore \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}
$$

Thus, we can show that the likelihood of a Gaussian sample also follows a Gaussian distribution. Reference