Joscandy Nunez
CPE 695

Q.1

Bayes' Theorem is useful for machine learning problems because it aids in predicting the probability of a hypothesis being true or not given some observed evidence. The theorem does not merely calculate probability but also takes into account the correlation and dependency between X and z. This theorem allows for the machine learning problems to be built upon numerical models of a machine's beliefs, therefore, being a good measure of performance for classifiers.

Equation 9-2. Bayes' theorem

$$p(\mathbf{z}\,|\,\mathbf{X}) = \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} = \frac{p(\mathbf{X}\,|\,\mathbf{z})\,p(\mathbf{z})}{p(\mathbf{X})}$$

Equation: Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow
by Aurélien Géron

Let's assume the probability of both X and z happening is P(X and z)
The probability that P(X and z) is true is the same as X being true, given z, multiplied by the probability of z being true:

P(X and z) = P(z) P(X | z)

Since P(X and z) can also be considered in terms of z being true, given X, we can also rewrite
P(X and z) = P(X) P(z | X)

Therefore,

P(X) P(z | X) = P(X and z) = P(z) P(X | z)

P(X) P(z | X) = P(z) P(X | z)

$$P(X \mid z) = \frac{P(z \mid X)\,P(X)}{P(z)}$$

or

$$P(z \mid X) = \frac{P(X \mid z)\,P(z)}{P(X)}$$

Joscandy Nunez
CPE 695

Q.2

Given the hypothesis function:

$$h_w(x) = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_n x_n$$

To minimize the cost function of the Ridge Regression:

$$E(w) = \sum_{i=1}^{m} (w^T \cdot x^{(i)} - y^{(i)})^2 + \lambda \sum_{i=1}^{m} w_i^2$$

Let's assume that, similar to linear regression, we can represent the problem in matrix notation. Input samples m is also a column vector.

For linear regression, we rewrite E(w) as $h_\theta(x) = \theta^T x$

For Ridge Regression, we add the penalty $\lambda \sum_{i=1}^{m} w_i^2$

This penalty, without x scalar multiplication, is $\lambda(\theta^T)^2$

Replacing the explicit sum by matrix multiplication, and carrying our added penalty in terms of θ :

$$J(\theta) = \frac{1}{2m} (X\theta - y)^T (X\theta - y) + \lambda(\theta^T)^2$$

Now, we can use matrix transpose identities, and also throw away the m part:

$$J(\theta) = ((X\theta)^T - y^T)(X\theta - y) + \lambda(\theta^T\theta^T)$$

$$J(\theta) = (X\theta)^T X\theta - (X\theta)^T y - y^T (X\theta) + y^T y + \lambda(\theta^T \theta^T)$$

Vector multiplication to simplify further:

$$J(\theta) = (\theta^T X^T X\theta - 2(X\theta)^T y + y^T y + \lambda(\theta^T \theta^T)$$

To find min, derive by θ:

$$\frac{dE}{d\theta} = 2X^T X\theta - 2X^T y + 2\lambda\theta$$

Joscandy Nunez
CPE 695

Compare to 0:

$$2X^T X\theta \; - \; 2X^T y \; + \; 2\lambda\theta = 0$$

$$2X^T X\theta \; + \; 2\lambda\theta \; = \; 2X^T y$$

2's drop:

$$X^T X\theta \; + \; \lambda\theta \; = \; X^T y$$

Isolate θ:

$$\theta \; = \; (X^T X \; + \; \lambda I)^{-1} \; X^T y$$

By this proof, the normal equation is indeed $w \; = \; (\lambda I \; + \; X^T \; \cdot X)^{-1} \cdot X^T \cdot y$

Reference: Derivation of the Normal Equation for linear regression
by Eli Bendersky

Joscandy Nunez
CPE 695

Q.3.1

We would need to estimate (n+1) * k parameters, where n+1 is the number of coefficients for each class, and k is the total number of classes.
These coefficients are a mapping between each feature and each class.

$$J(\theta) \;=\; -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} y_k\ (i)\ log\ (\widehat{p}_k(i))$$

Derivative of $J(\theta)$:

Substitute in softmax:

$$p_k(i) \;=\; softmax(z_k(i)) = (\frac{exp(s_k(x))}{\sum_{j=1}^{K} exp((s_j(x))})$$

$$J(\theta) \;=\; -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} y_k\ (i)\ log\ (\frac{exp(s_k(x))}{\sum_{j=1}^{K} exp((s_j(x))})$$

Compute the partial derivative of $J(\theta)$:

y = $\hat{y} = h_w(x) = w^T * x$

derivative of above y' = $x^{(i)}$

Derivative of ln(x) = $\frac{1}{x}$, with chain rule:

$$= \;-\frac{1}{m}\sum_{i=1}^{m} x^{(i)} \;-\; \frac{1}{exp(s_j(x)} \;*\; exp(s_k(x))\, x^{(i)}$$

$$s_k(x) \;=\; X^T\theta^{(k)}$$

sigma is placed back as weights to follow y = $\hat{y} = h_w(x) = w^T * x$

$$= \;-\frac{1}{m}\sum_{i=1}^{m} x^{(i)} \;-\; \frac{1}{exp(\,X^T\theta^{(j)})} \;*\; exp(\,X^T\theta^{(k)})\,\sum x^{(i)}$$

$$= \frac{1}{m}\sum_{i=1}^{m} x^{(i)}\ (\widehat{p}_k(i) \;-\; y^{(i)}{}_k)$$