Erin Brown
02/22/2023

1. **[*6 points*] Prove Bayes' Theorem. Briefly explain why it is useful for machine learning problems**.

Bayes' Theorem is the probability of two events, A and B, occurring when there is nonzero probability of occurrence. It is useful for machine learning because as new data comes in, it will update probabilities to obtain more accurate results.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Proof by Derivation

The rule for conditional probability is,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

This can be written reversally as,

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

The numerators of both equations P(A|B) and P(B|A) are the same because the probability of A and B occurring is the same as the probability of B and A occurring. Therefore, we can rewrite our reverse equation as,

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

Solve for P(A ∩ B)

$$P(B|A) \times P(A) = \frac{P(A \cap B)}{P(A)} \times P(A)$$

Thus

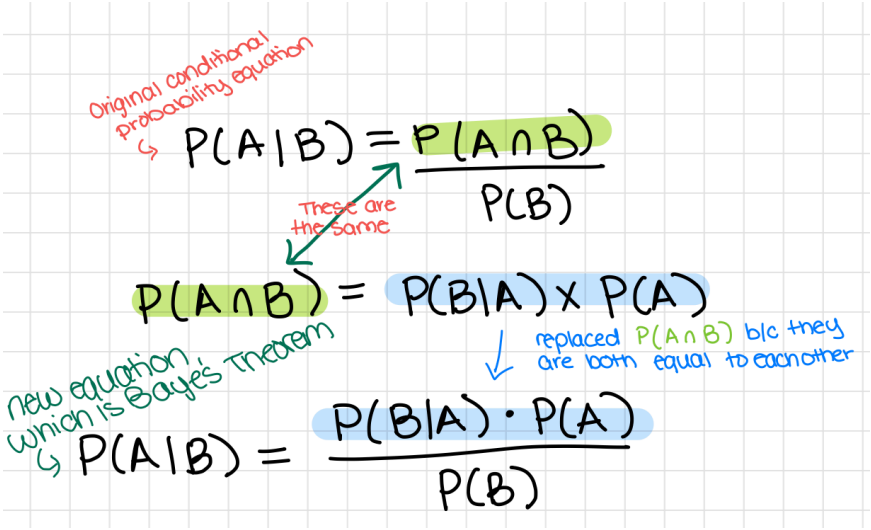
$$P(A \cap B) = P(B|A) \times P(A)$$

Referring back to our rule for conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

We see that P(A ∩ B) in the numerator can now be replaced by our solution P(B|A) x P(A). This will give,

Erin Brown
02/22/2023



As shown above, the derivation gives us Bayes' Theorem,

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

2. [10 points] In Module 2, we gave the normal equation (i.e., closed-form solution) for linear regression using MSE as the cost function. **Prove that the closed-form solution for Ridge Regression** is $\boldsymbol{w} = (\lambda I + X^T \cdot X)^{-1} \cdot X^T \cdot \boldsymbol{y}$, where I is the identity matrix, $X = (x^{(1)}, x^{(2)}, \dots, x^{(m)})^T$ is the input data matrix, $x^{(i)} = (1, x_1, x_2, \dots, x_n)$ is the i-th data sample, and $\boldsymbol{y} = (y^{(1)}, y^{(2)}, \dots, y^m)$. Assume the hypothesis function $h_w(x) = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$, and $y^{(j)}$ is the measurement of

$h_w(x)$ for the j-th training sample. The cost function of the Ridge Regression is $E(\boldsymbol{w}) = \sum_{i=1}^{m}(\boldsymbol{w}^T \cdot \boldsymbol{x}^{(i)} - y^{(i)})^2 + \lambda \sum_{i=1}^{m} w_{i2}$.

Erin Brown
02/22/2023

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

From normal form

$$J(w) = (xw-y)^T(xw-y)$$

$$(xw-y)^T(xw-y) + \lambda w^T w$$

$$\frac{\delta (xw-y)^T(xw-y)}{\delta w} = -2x(y - w^T x)$$

and

$$\frac{\delta (\lambda w^T w)}{\delta w} = 2\lambda w$$

so,

$$x^T y = x^T x w + \lambda w$$

Derived for w

$$x^T(xw) = (xw)^T x = w^T x^T w$$

$$x^T y = w^T x^T x + \lambda w$$

$$w^T x^T x + \lambda w - x^T y = 0$$

$$w^T(x^T x + \lambda I) - x^T y = 0$$

$$w^T(x^T x + \lambda I) = x^T y$$

$$w = (x^T x + \lambda I)^{-1} x^T y$$

3. [10 points] Assume we have K different classes in a multi-class Softmax Regression model. The posterior probability is $\hat{p}_k = \delta(s_k(x))_k = \frac{\exp(s_k(x))}{\sum_{j=1}^{K} \exp(s_j(x))}$ for $k = 1, 2, \ldots, K$, where $s_k(x) = \theta_k{}^T . x$, input x is an n-dimension vector, and K the total number of classes.
   1) To learn this Softmax Regression model, how many parameters we need to estimate? What are these parameters?

n by k, These parameters would be

$$\Theta = \begin{bmatrix} | & | & | & & | \\ \theta^{(1)} & \theta^{(2)} & \theta^{(3)} & \cdots & \theta^{(K)} \\ | & | & | & & | \end{bmatrix}$$

Erin Brown
02/22/2023

2) Consider the cross-entropy cost function $J(\Theta)$ of m training samples $\{(x_i, y_i)\}_{i=1,2,\ldots,m}$ as below. Derive the gradient of $J(\Theta)$ regarding to θ_k.

$$J(\Theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log\left(p_k^{(i)}\right)$$

where $y_k(i) = 1$ if the i[th] instance belongs to class k; 0 otherwise.

$$\nabla J(\Theta) = 1/m \sum_{i=1}^{m} (\hat{p}_k^{(i)} - y_k^{(i)}) x^{(i)}$$

$$\hat{p}_k = \delta(s_k(x))_k = \frac{e^{x(\theta_k^T \cdot x)}}{\sum_{j=1}^{k} e^{x(\theta_j^T \cdot x)}}$$

$$J\Theta = -1/m \sum_{i=1}^{m}\sum_{k=1}^{k} y_k^{(i)} \log\left(\frac{e^{x(\theta_k^T \cdot x^{(i)}}}{\sum_{j=1}^{k} e^{x(\theta_j^T \cdot x^{(i)})}}\right)$$

$$= -1/m \sum_{i=1}^{m}\sum_{k=1}^{k} y_k^{(i)} [\log(e^{x\theta_j^T \cdot x^{(i)}})) - \log(\sum_{j=1}^{k} e^{x(\theta_j^T \cdot x^{(i)})})]$$

$$= -1/m \left(\sum_{i=1}^{m}\sum_{k=1}^{k} y_k^{(i)} \log(e^{x(\theta_j^T \cdot x^{(i)})}) + \sum_{i=1}^{m}\sum_{k=1}^{k} y_k^{(i)} \log\left(\sum_{j=1}^{k} e^{x(\theta_j^T \cdot x^{(i)})}\right)\right)$$

$$= -1/m \left(\sum_{i=1}^{m}\sum_{k=1}^{k} y_k^{(i)} \cdot \theta_j^T \cdot x^{(i)} + \sum_{i=1}^{m}\sum_{k=1}^{k} y_k^{(i)} \log\left(\sum_{j=1}^{k} e^{x(\theta_j^T \cdot x^{(i)})}\right)\right)$$

$$\nabla J(\theta) = 1/m \sum_{i=1}^{m} y_k^{(i)} \cdot x + \sum_{i=1}^{m} \frac{1}{\sum_{j=1}^{k} e^{x(\theta_j^T \cdot x)}} e^{x(\theta_k^T x^{(i)})} \cdot x^{(i)}$$

$$\nabla_{\theta_k} J(\theta) = 1/m \sum_{i=1}^{m} \left(\frac{e^{x(\theta_k^T \cdot x)}}{\sum_{j=1}^{k} e^{x(\theta_j^T \cdot x)}} - y_k^{(i)}\right) x^{(i)}$$

$$\hat{p}_k = \delta(s_k(x))_k = \frac{e^{x(\theta_k^T \cdot x)}}{\sum_{j=1}^{k} e^{x(\theta_j^T \cdot x)}}$$

$$\nabla_{\theta_k} J(\theta) = 1/m \sum_{i=1}^{m} (\hat{p}_k - y_k^{(i)}) x^{(i)}$$