CS 529: Homework 1

Andrea Zapata - PUID : 0031827996

1. (a)
   - Generalization : being able to correctly categorize new input vectors that are not present in the training data.
   - Overfitting : having a model that performs too well on training data but does not perform well on new input vectors.
   - Underfitting : having a model which is too simple to represent the underlying model of the training data and will not perform well on new input vectors.
   - Regularization : including a parameter in the error function that controls the complexity of the model in order to avoid overfitting.
   - No free lunch theorem : this theorem states that if a models A performs better than model B on a data-set X, there will be a data-set Y where model B performs better than A.
   - Occams razor: giving many models who can achieve the same test error, one should choose the least complex model.
   - Independent and identically distributed data points: having independent and identically distributed data points means that all data points come from the same distribution and were sampled individually.
   - Cross-validation: dividing the training data in k groups, then use (k-1) groups for training and select one group for testing the model. This is a strategy to predict the performance of the model on test data.
   - Degrees of freedom: the number of independent variables on which the model depends.

   (b) For $Coin_1$, $P(x = H) = 12/17$ and $P(x = T) = 5/17$ and for $Coin_2$, $P(x = H) = 6/17$ and $P(x = T) = 11/17$. As the coin tosses are independent random variables, the order of the results does not matter, so the correct guess for $Coin_1$ would be "H" and for $Coin_2$ would be "T" as each value has a higher probability within what has been observed about their distribution.

   (c) Let's define :

$$\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_n \end{bmatrix} = \begin{bmatrix} t_1 \\ t_2 \\ ... \\ t_n \end{bmatrix} - \begin{bmatrix} 1 & x_{11} & x_{21} & ... & x_{m1} \\ 1 & x_{12} & x_{12} & ... & x_{m2} \\ ... & ... & ... & ... & ... \\ 1 & x_{1n} & x_{2n} & ... & x_{mn} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ ... \\ w_m \end{bmatrix} \tag{1}$$

then $\epsilon = t - \mathbf{X}w$, we could express then $E(w)$ as:

$$
\begin{aligned}
E(w) &= \frac{1}{2}\begin{bmatrix} \epsilon_1 & \epsilon_2 & ... & \epsilon_n \end{bmatrix}\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_n \end{bmatrix} = \frac{1}{2}\left[\epsilon_1\epsilon_1 + \epsilon_2\epsilon_2 + ... + \epsilon_n\epsilon_n\right] = \frac{1}{2}\epsilon^T\epsilon \\
&= \frac{1}{2}(t - \mathbf{X}w)^T(t - \mathbf{X}w) = \frac{1}{2}(t^T - w^T\mathbf{X}^T)(t - \mathbf{X}w) \\
&= \frac{1}{2}\left[t^T t - t^T\mathbf{X}w - w^T\mathbf{X}^T t + w^T\mathbf{X}^T\mathbf{X}w\right] \\
&= \frac{1}{2}t^T t - w^T\mathbf{X}^T t + \frac{1}{2}w^T\mathbf{X}^T\mathbf{X}w
\end{aligned}
$$

To minimize the error we set the derivative to zero :

$$
\begin{aligned}
\frac{\delta E(w)}{\delta w} &= \frac{\delta}{\delta w}\left[\frac{1}{2}t^T t - w^T\mathbf{X}^T t + \frac{1}{2}w^T\mathbf{X}^T\mathbf{X}w\right] \\
&= -\mathbf{X}^T t + \frac{1}{2}\mathbf{X}^T\mathbf{X}w + \frac{1}{2}\mathbf{X}^T\mathbf{X}w \\
&= -\mathbf{X}^T t + \mathbf{X}^T\mathbf{X}w = 0
\end{aligned}
$$

then

$$
\begin{aligned}
\mathbf{X}^T\mathbf{X}w &= \mathbf{X}^T t \\
w^* &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T t
\end{aligned}
$$

Finally, we can represent $y(x, w^*)$ as

$$
y(x, w^*) = x^T w^* = x^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T t
$$