ENGN2520 Midterm

Ming Xu (Banner ID: B01532164)

Problem1

(a) The classification rules defined by the Bayes optimal classifier

Regarding the length and weight of the given fish:

$If\ P(length|salmon) * P(weight|salmon) * P(salmon) > P(length|bass) * P(weight|bass) * P(bass)$, then $f(x) = salmon$. Otherwise, $f(x) = bass$.

(b) Possible reasons:

1. The length and weight of a type of fish are not distributed normally.
2. The length and weight of a given fish is highly corelated. It's wrong to just multiply the two possibilities.
3. The training set is not big enough. Or the training set may have bias and give little useful information.

Problem2

(a) The degree polynomial shouldn't be too big to avoid overfitting, for example, the degree should not be greater than the number of training data. Typically, we can partition the training data into a training set and a validation set. After we train the models of different degree on the training set, we can validate the regression model on the validation set to pick up the best degree.

(b) Multilayer neural network has the ability to learn and model very complicated non-linear models.

(c) Linear classifier is easy to setup and model. Also, multilayer neural network may be sensitive to small noise.

(d) I suppose we should pick c_2. Since the VC dimension of H2 is bigger, c_2 is less likely to overfit. Also, c_2 utilizes more features of a data point, it seems more convincing.

Problem3

$$\boldsymbol{\phi}(\boldsymbol{x}) = \begin{bmatrix} |\boldsymbol{x} - \boldsymbol{a}| \\ |\boldsymbol{x} - \boldsymbol{b}| \end{bmatrix}, \boldsymbol{w} = \begin{bmatrix} \mathbf{1} \\ \mathbf{1} \end{bmatrix}$$

The intuition is that if a point x falls into an interval, then the distance to the two boundaries sum up to the length of the interval. If a point x falls outside the interval, then the sum of the distance to two boundaries must be larger than the length of the interval.

Therefore, the classifier C: $X \to \{-1, +1\}$ is:

$$\boldsymbol{C}(\boldsymbol{x}) = \begin{cases} +\mathbf{1} & \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\phi}(\boldsymbol{x}) \leq |\boldsymbol{b} - \boldsymbol{a}| \\ -\mathbf{1} & \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\phi}(\boldsymbol{x}) > |\boldsymbol{b} - \boldsymbol{a}| \end{cases}$$

Problem4

(a) The probability density of x is:

$$\boldsymbol{p_1(x) * w_1 + p_2(x) * w_2}$$

(b) The algorithm to calculate $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$:

1. Set the initial value of $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ by 0.5 and 0.5, respectively.

2. In each iteration, traverse all data point to compare:

$$\boldsymbol{p_1(x) * w_1 \; and \; p_2(x) * w_2}$$

to classify the current data and record its specie.

3. After all data are traversed, count the number of tigers and cheetahs, and update the $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ accordingly.

4. Repeat the above iteration(Step2, 3) until the $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ are stable(the absolute difference between current $\boldsymbol{w_1}$ and the $\boldsymbol{w_1}$ from previous iteration smaller than a threshold).