Machine Learning
Homework 1

Due on March 22, 2017

1. Let

$$A_+ = \{(0,0)\ (1,1),\ (-1,1),\ (1,-1),\ (-1,-1)\}$$

and

$$A_- = \{(1,0),\ (-1,0),\ (0,1),\ (0,-1)\}$$

represent the *positive* and *negative* training instances respectively.

(a) Plot the decision boundary for the 3-nearest neighbor algorithm.

(b) What is the training set accuaracy for 3-nearest neighbor algorithm?

(c) What is the *confusion* matrix for 3-nearest neighbor algorithm on the training set?

2. Let the $A \in R^{n \times n}$ be a symmetric positive definite matrix. Show that all eigenvalues of matrix A are positive.

3. Let $Z = [X_1; X_2; X_3]$ be a random vector and Σ be a matrix with size 3×3 where $\Sigma_{ij} = Cov(X_i, X_j)$ and $\Sigma_{ii} = Var(X_i)$. Let the random variable $W = \mathbf{a}^\top Z = a_1X_1 + a_2X_2 + a_3X_3$ where $\mathbf{a} = [a_1; a_2; a_3]$. *i.e.*, the random variable W is the *projection* of random vector Z onto the vector $\mathbf{a}$. Find the variance of W.

4. Let S be a set of $10,000$ random numbers generated by the *Gaussian distribution*, $\mathcal{N}(0;1)$. You have to estimate the *mean* and *standard deviation* of this *Gaussian distribution*.

   (a) Randomly select 10 number from the set S and then use the average of these 10 number as the estimation. Repeat this experiment 20 times. What is the *sample mean* and *sample standard deviation* of these 20 random experiments?

   (b) Randomly select $1,000$ number from the set S and then use the average of these $1,000$ number as the estimation. Repeat this experiment 50 times. What is the *sample mean* and *sample standard deviation* of these 50 random experiments?

5. Consider a linear system of equations as follows:

$$\begin{array}{rcrcrcl} 2x_1 & + & 2x_2 & - & x_3 & = & 8 \\ x_1 & - & x_2 & + & 2x_3 & = & 0 \\ x_1 & + & 2x_2 & - & 2x_3 & = & 4 \\ 4x_1 & - & x_2 & - & 3x_3 & = & 0 \end{array}$$

   Find the least squares approximation solution for it.

6. Generate a training dataset with size $1,000$ by yourself.

   (a) That is,

$$S = \{(\mathbf{x}^i, y_i) | \mathbf{x}^i = (\mathbf{x}_1^i,\ \mathbf{x}_2^i) \in R^2, \text{and } y_i \in R, i = 1, \ldots, 1,000\},$$

   where $\mathbf{x}_1$ and $\mathbf{x}_2$ are generated by the *uniform distribution*, $U[-1;1]$. The observation value $y = 2\mathbf{x}_1^2 + \mathbf{x}_2^2 + 1.5\mathbf{x}_1\mathbf{x}_2 + \mathbf{x}_1 + 2\mathbf{x}_2 + \epsilon$ where ϵ is the random noise generated by $\mathcal{N}(0,1)$.

   (b) Find a *quadratic* function $f(\mathbf{x})$, that is fitted in the training dataset S.

   (c) Compute the MAE, mean of absoulte error, and plot the function you get.