Assignment 2

Due: 2015/11/11 23:59:59

1. Given a training dataset $\mathcal{X} = \{x^{(t)}, \boldsymbol{r}^{(t)}\}_{t=1}^{N}$ where $x^{(t)} \in \mathbb{R}$ are scalars and the number of classes is $K = 2$. Suppose instances are normally distributed within each class. Write in closes-form the decision boundary $z \in \mathbb{R}$, where $P[C_1|z] = P[C_2|z]$.

2. Consider the univariate parametric classification. Show that the priors $P[C_i]$, $i = 1, 2, \cdots, K$, for different classes can be estimated jointly by assuming that $P[C_i]$ follows a Multinomial distribution parametrized by $\theta = (\rho_1, \cdots, \rho_K)$ with constrains $\sum_{i=1}^{K} \rho_i = 1$ and by maximizing the likelihood $P[\mathcal{X}|\theta]$.

3. Show that the *Area Under the ROC Curve* (AUC) is equal to the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative one.
(Hint: by partitioning the AUC horizontally)

4. [Coding Assignment] Train a classifier using the probabilistic models such that it is able to predict labels given new examples. You are free to apply any data preprocessing technique, model selection technique, and ensemble method you have learned to improve the prediction accuracy. See coding spec for more details.