# CS57800 Statistical Machine Learning
## Homework 3

**Ting Zhang**
School of Industiral Engineering
zhan1013@purdue.edu

November 6, 2015

## 1 Foundations

1. The VC dimension of C is $2n + 1$.

2. To show it is a convex function, we can compute the second derivative of the function.

$$\frac{dl}{d\hat{y}} = \frac{1}{\log 2}[\frac{-ye^{-y\hat{y}}}{(1+e^{-y\hat{y}})\log 2}] = \frac{1}{(\log 2)^2}(-1 + \frac{y}{1+e^{-y\hat{y}}})$$
$$\frac{d^2l}{d\hat{y}^2} = \frac{1}{(\log 2)^2}[0 + y * (-2) * (\frac{1}{1+e^{-y\hat{y}}})^2 * (-y) * e^{-y\hat{y}}] = \frac{2y^2e^{-y\hat{y}}}{(\log 2)^2(1+e^{-y\hat{y}})} > 0$$

Since the second derivative of this function is greater than 0, then this function is a convex function.

3. The training error of the final hypothesis can be formulated as

$$TrainingError(H_{final}) \leq \prod_t [2\sqrt{\epsilon_t(1-\epsilon_t)}].$$

4. The decision boundary of each weak hypothesis is a axis-parallel line, as shown in Figure **??**. Positive points are represented with "+", and negative points are represented with "*". Then the process of doing Adaboost is as following:
(1) First iteration $t = 1$, find the decision boundary that makes the least number of mistakes. In first iteration, each example is uniformlly distributed, with $D_1(i) = 0.1$. Here it found $x_1 > 6$, with 2 mistakes, including example 9 and 10. Therefore, the error is $\epsilon_1 = 0.2$. The weak hypothesis is $h_1(x_i) : sign(x_1 - 6)$.
(2) Second iteration $t = 2$, first, update the distribution of each example. To update the distribution, we need to follow these euqations:

$$\alpha_1 = 0.5 * \ln(\frac{1-\epsilon_1}{\epsilon_1}) = 0.5 * \ln(\frac{1-0.2}{0.2}) = 0.693e_{-\alpha_1} = 0.500e_{\alpha_1} = 2.000$$

So, the distribution of the correctly classified examples $i = 1, 2, 3, 4, 5, 6, 7, 8$ becomes

$$D_2(i) = \frac{D_1(i) * e_{-\alpha_1}}{Z_1} = \frac{0.05}{Z_1}$$

and the distribution of the wrongly classified examples $i = 9, 10$ becomes

$$D_2(i) = \frac{D_1(i) * e_{\alpha_1}}{Z_1} = \frac{0.2}{Z_1}$$

Since $\sum_{i=1}^{10} D_2(i) = 1$, we have $(8 * 0.05 + 2 * 0.2)/Z_1 = 1$. So, $Z_1 = 0.8$. Therefore, the distribution of the examples now become

$$D_2(i) = \begin{cases} 0.0625 & \text{if } i = 1, 2, 3, 4, 5, 6, 7, 8 \\ 0.25 & \text{if } i = 9, 10 \end{cases}$$

Then, finding the weak hypothesis that has the smallest error $\sum_{i=1} D_2(i)$, where $i$ is the index of wrongly classified example. Then, it find the weak hypothesis to be $h_2(x_i) : sign(x_2 - 8)$, with four mistakes, including example 3, 6, 7, 9. And the error is $\epsilon_2 = 0.25$. We can also compute $\alpha_2 = 0.5 * \ln(\frac{1-\epsilon_2}{\epsilon_2}) = 0.549$.
Then, the final hypothesis is

$$H_{final}(x) = sign(\sum_t \alpha_t h_t(x)) = sign(0.693 * sign(x_1 - 6) + 0.549 * sign(x_2 - 8))$$

5. To verify a kernel function, we need to proof the positive semi-definite property of it.

$$\begin{aligned}
&\iint f(\vec{x})K(\vec{x}, \vec{y})f(\vec{y})d\vec{x}d\vec{y} \\
&= \iint f(\vec{x})[\alpha K_1(\vec{x}, \vec{y}) + \beta K_2(\vec{x}, \vec{y})]f(\vec{y})d\vec{x}d\vec{y} \\
&= \iint \alpha f(\vec{x})K_1(\vec{x}, \vec{y})f(\vec{y})d\vec{x}d\vec{y} + \beta f(\vec{x})K_2(\vec{x}, \vec{y})f(\vec{y})d\vec{x}d\vec{y} \\
&> 0 * 0 + 0 * 0 \\
&= 0
\end{aligned}$$

Because, $K_1(\vec{x}, \vec{y})$ and $K_2(\vec{x}, \vec{y})$ are positive semi-definite, $\alpha$ and $\beta$ are all positive.

6. $\xi$ is defined as the margin of the soft SVM. When an example is classified correctly, $0 \leq \xi \leq 1$; while if an example is classified wrongly, then $\xi > 1$. Therefore, with $M$ examples, there are at most $M$ mistakes. Since $\xi > 1$, then $\sum_{i=1}^{M} \xi_i > M * 1 = M$. Therefore, $\sum_{i=1}^{M} \xi_i$ is the upper bound on the training error of the classifier.

# 2 Programming Report

...