Machine Learning & Pattern Recognition

SONG Xuemeng

songxuemeng@sdu.edu.cn

http://xuemeng.bitcron.com/

1. **Ture of False: 17*2 pts**
2. **Short Questions: 6*4 pts**
3. **Long Questions: 4, 42 pts in total**

   A. B. C. D.

   May be Loss Function, Gradient Decent, SVM, Decision Tree, Boosting, PCA, Kernel, LDA, GMM, Neural Network…

1. Ture of False: 17*2 pts

- (T/F) Decision tree is learned by minimizing information gain.
- (T/F) The VC dimension of a line should be at most 2, since I can find at least one case of 3 points that cannot be shattered by any line.
- (T/F) GMM is a universal approximator of densities, i.e., an arbitrary density f(·), can be approximated by a Gaussian mixture model.
- (T/F) l_1 norm shrinks parameter values towards zero (parameter shrinkage).

2. Short Questions: 6*4 pts

- What is the similarity and difference between the PCA and LDA? (**4 points)**
- How to alleviate the overfitting problem? Name the possible actions. **(4 points)**
- What is the VC-dimension of the SVM with Gaussian Kernel? Explain your answer. Do not need to provide the detailed proof. **(4 points)**

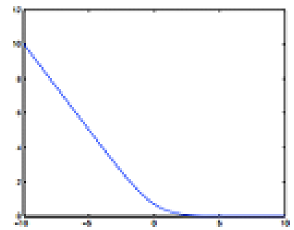
3. Long Questions: 4, 42 pts in total

3. Long Questions:

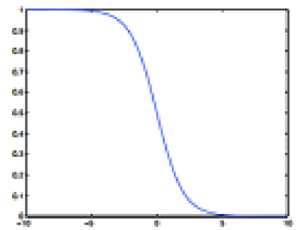
B. Loss Function (12 points) Generally speaking, a classifier can be written as $H(x) = sign(F(x))$, where $H(x)$: $\mathbb{R}^d \rightarrow \{-1, 1\}$ and $F(x)$: $\mathbb{R}^d \rightarrow \mathbb{R}$. To obtain the parameters in $F(x)$ we need to minimize the loss function averaged over the training set: $\sum_i L(y^i F(x^i))$. Here L is a function of $yF(x)$. For example, for linear classifiers,

$$F(x) = w_0 + \sum_{j=1}^{d} w_j x_j \text{ , and } yF(x) = y(w_0 + \sum_{j=1}^{d} w_j x_j)$$

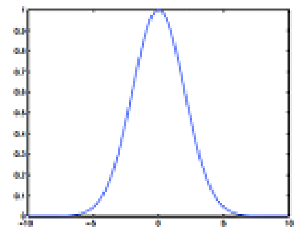
1. Which loss functions below are appropriate to use in classification? For the ones that are not appropriate, explain why not. In general, what conditions does L have to satisfy in order to be an appropriate loss function? The x axis is $yF(x)$, and the y axis is $L(yF(x))$. **(4 points)**



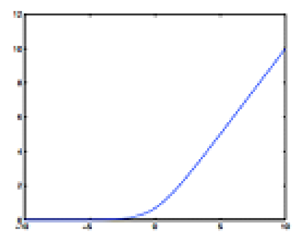
(a)



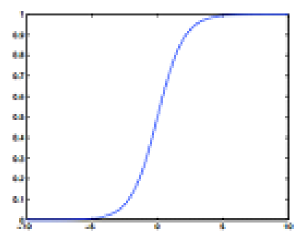
(b)



(c)



(d)



(e)

2. Of the above loss functions appropriate to use in classification, which one is the most robust to outliers? Justify your answer. **(4 points)**

3. Let $F(x) = w_0 + \sum_{j=1}^{d} w_j x_j$ and $L(yF(x)) = \frac{1}{1+\exp(yF(x))}$. Suppose you use gradient descent to obtain the optimal parameters w_0 and w_i. Give the update rules for these parameters. **(4 points)**