Lec Mon

Monday, November 19, 2018 4:05 PM

- Soft SVM allows for some slack to accommodate data sets that aren't linearly separable

$$\min \frac{1}{2} w^T w + C \sum_i \xi_i$$

if C is 0, the minimization becomes meaningless
if C is infinity, we have hard SVM

equivalently,

$$\min \frac{1}{2} w^T w + C \sum_i \max(0, 1 - y_i(w^T x_i + b))$$

C is a hyper parameter controlling tradeoff between large margin and small hinge-loss

How do we find the minimum? The hinge loss function is not differentiable because there's a sharp corner

We can use stochastic sub-gradient descent (sub-gradients not in exam)
I
We can map data to a very high dimensional space to make data separable

Multi-Class classification
In the real world, our labels will often be more than binary

Two key ideas to solve multiclass
- reduce multiclass to binary
    - One Against All - decompose multiclass prediction into multiple binary decisions
        - eg for labels {1, 2, 3, …, K} we train K different models, each telling us whether or not the example has the nth label
    - One vs. One - always pick two different classes and compare
        - there are k choose 2 pairs of classes
        - we can make a "decision tree" of comparisons, height K-1 so we have to make that many binary classifications

Comparison
- One against all
    - O(K) weight vectors to train and store
    - training set of binary classifiers may be unbalanced
    - less expressive, make a strong assumption
- One vs. One
    - O(K^2) weight vectors to train and store
    - size of training set for a pair of labels could be small
        - overfitting of binary classifiers
    - need large space to store model