CS489/698
Lecture 14: February 27, 2017

Support Vector Machines (continued)

[B] Sec. 7.1 [D] Sec. 11.5-11.6 [HTF] Chap. 12 [M] Sec. 14.5 [RN] 18.9 [MRT] Chap. 4

Overlapping Class Distributions

- So far we assumed that data is linearly separable
  - High dimensions help for linear separability, but may hurt for generalization
- But what if the data is noisy (mistakes or outliers)
  - Constraints should allow misclassifications
- Picture

Soft margin

- Idea: relax constraints by introducing slack variables $\xi_n \geq 0$

$$y_n \, \boldsymbol{w}^{\boldsymbol{T}} \phi(\boldsymbol{x}_{\boldsymbol{n}}) \geq 1 - \xi_n \quad \forall n$$

- Picture:

Soft margin classifier

- New optimization problem:

$$\min_{\boldsymbol{w},\boldsymbol{\xi}} \quad C\sum_{n=1}^{N}\xi_n + \frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2$$

s.t. $y_n\,\boldsymbol{w}^{\boldsymbol{T}}\phi(\boldsymbol{x}_{\boldsymbol{n}}) \geq 1 - \xi_n$

and $\xi_n \geq 0 \quad \forall n$

- where $C > 0$ controls the trade-off between the slack variable penalty and the margin

Soft margin classifier

- Notes:
  1. Since $\sum_n \xi_n$ is an upper bound on the # of misclassifications, C can also be thought as a regularization coefficient that controls the trade-off between error minimization and model complexity
  2. When $C \to \infty$, then we recover the original hard margin classifier
  3. Soft margins handle minor misclassifications, but the classifier is still very sensitive to outliers

Support Vectors

- As before support vectors correspond to active constraints

$$y_n \, \boldsymbol{w}^{\boldsymbol{T}} \phi(\boldsymbol{x}_{\boldsymbol{n}}) = 1 - \xi_n$$

  - i.e., all points that are in the margin or misclassified
- Picture:

Multiclass SVMs

- Three methods:
  1. One-against-all: for K classes, train K SVMs to distinguish each class from the rest
  2. Pairwise comparison: train $O(K^2)$ SVMs to compare each pair of classes
  3. Continuous ranking: single SVM that returns a continuous value to rank all classes

One-Against-All

- For K classes, train K SVMs to distinguish each class from the rest
- Picture:

- Problem: what if different classes are returned by different SVMs?

Pairwise Comparison

- Train $O(K^2)$ SVMs to compare each pair of classes
- Picture:

- Problem: how do we pick the best class?

Continuous Ranking

- Single SVM that returns a continuous value to rank all classes
- Picture:

- Most popular approach today

Continuous Ranking

- Idea: instead of computing the sign of a linear separator, compare the values of linear functions for each class k

- Classification:

$$y_* = argmax_k \, \boldsymbol{w}_k^T \phi(\boldsymbol{x}_*)$$

Multiclass Margin

- For each class $k \neq y$ define a linear constraint:

$$\boldsymbol{w}_y^T \phi(\boldsymbol{x}) - \boldsymbol{w}_k^T \phi(\boldsymbol{x}) \geq 1 \quad \forall k \neq y$$

- This guarantees a margin of at least 1

Multiclass Classification

- Optimization problem:

$$\min_{\boldsymbol{W}} \frac{1}{2}\sum_k ||\boldsymbol{w}_k||^2$$

$$\text{s.t.} \quad \boldsymbol{w}_{y_n}^T \phi(\boldsymbol{x}_n) - \boldsymbol{w}_k^T \phi(\boldsymbol{x}_n) \geq 1 \quad \forall n, k \neq y_n$$

- Equivalent to binary SVM when we have only two classes

Overlapping classes

- Add slack variables:

$$\min_{\boldsymbol{W},\boldsymbol{\xi}} \; C\sum_n \xi_n + \frac{1}{2}\sum_k \left|\left|\boldsymbol{w}_k\right|\right|^2$$

$$\text{s.t. } \mathbf{w}_{y_n}^T \phi(\boldsymbol{x}_n) - \boldsymbol{w}_k^T \phi(\boldsymbol{x}_n) \geq 1 - \xi_n \quad \forall n, k \neq y_k$$

- Equivalent to binary SVM when we have only two classes