Lecture 24: Support Vector Machines
CS480/680 Intro to Machine Learning

2023-4-6

Pascal Poupart
David R. Cheriton School of Computer Science

UNIVERSITY OF WATERLOO

Sparse kernel techniques

- Kernel based approaches: complexity depends on the amount of data, not the dimensionality of the space. But for large datasets, this is not practical.
  - Kernel matrix is square in # of data points
  - Prediction requires inversion of the kernel matrix, which is cubic in # of data points

- Can we use a **sparse representation**?
  - i.e., kernel that depends on a subset of the data

Support Vector Machines

- Kernel depends on subset of data
- Picture

Max-Margin Classifier

- Find linear separator that maximizes the distance (or margin) to closest data points
- Picture

Margin

- Linear separator: $\boldsymbol{w}^T\phi(\boldsymbol{x}) = 0$

- Distance to linear separator:

$$\frac{y\boldsymbol{w}^T\phi(\boldsymbol{x})}{||\boldsymbol{w}||} \quad \text{where } y \in \{-1,1\}$$

- Maximum margin:

$$max_{\boldsymbol{w}} \frac{1}{||\boldsymbol{w}||}\left\{\min_n y_n\, \boldsymbol{w}^{\boldsymbol{T}}\phi(\boldsymbol{x}_{\boldsymbol{n}})\right\}$$

Comparison

Perceptron

Support Vector Machine

Maximum Margin

- Unique max margin linear separator

$$max_{\boldsymbol{w}} \frac{1}{||\boldsymbol{w}||} \left\{ \min_n y_n \, \boldsymbol{w}^T \phi(\boldsymbol{x}_n) \right\}$$

- Alternatively, we can fix the minimal distance to 1 and minimize $||\boldsymbol{w}||$

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

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

- This is a convex quadratic optimization problem that can easily be solved by many optimization packages

Derivation

$$argmax_{\boldsymbol{w}} \frac{1}{||\boldsymbol{w}||} \left\{ \min_n y_n\, \boldsymbol{w}^T \phi(\boldsymbol{x}_n) \right\}$$

$$= argmax_{\boldsymbol{w},\delta} \frac{1}{||\boldsymbol{w}||} \delta \quad \text{s.t. } y_n\, \boldsymbol{w}^T \phi(\boldsymbol{x}_n) \geq \delta \;\; \forall n$$

$$= argmax_{\boldsymbol{w},\delta} \frac{1}{\left|\left|\frac{\boldsymbol{w}}{\delta}\right|\right|} \quad \text{s.t. } y_n \frac{\boldsymbol{w}^T}{\delta} \phi(\boldsymbol{x}_n) \geq 1 \;\; \forall n$$

replace $\frac{\boldsymbol{w}}{\delta}$ by $\boldsymbol{w}'$

$$= argmax_{\boldsymbol{w}'} \frac{1}{||\boldsymbol{w}'||} \quad \text{s.t. } y_n \boldsymbol{w}'^T \phi(\boldsymbol{x}_n) \geq 1 \;\; \forall n$$

$$= argmin_{\boldsymbol{w}'} ||\boldsymbol{w}'|| \quad \text{s.t. } y_n \boldsymbol{w}'^T \phi(\boldsymbol{x}_n) \geq 1 \;\; \forall n$$

$$= argmin_{\boldsymbol{w}'} \frac{1}{2} ||\boldsymbol{w}'||^2 \quad \text{s.t. } y_n \boldsymbol{w}'^T \phi(\boldsymbol{x}_n) \geq 1 \;\; \forall n$$

Support Vectors

- Quadratic optimization problem

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

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

- Only the points where $y_n \, \boldsymbol{w}^T \phi(\boldsymbol{x}_n) = 1$ are necessary. These points define the active constraints and are known as the **support vectors.**

Dual representation

- Idea: reformulation where $\phi(\boldsymbol{x})$ appears only in a kernel

- Approach: find the dual of the optimization problem

- Result: (sparse) kernel support vector machines

Dual derivation

- Transform constrained optimization

$$\min_{\boldsymbol{w}} \frac{1}{2}||\boldsymbol{w}||^2 \quad \text{s.t. } y_n\, \boldsymbol{w}^T \phi(\boldsymbol{x_n}) \geq 1 \quad \forall n$$

into an unconstrained optimization problem

- Lagrangian

$$\max_{\boldsymbol{a} \geq \boldsymbol{0}} \min_{\boldsymbol{w}} L(\boldsymbol{w}, \boldsymbol{a})$$

where $L(\boldsymbol{w}, \boldsymbol{a}) = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_n \underbrace{a_n [y_n\, \boldsymbol{w}^T \phi(\boldsymbol{x_n}) - 1]}_{\text{penalty for violating the } n^{\text{th}} \text{ constraint}}$

Dual derivation

- Solve inner minimization:

$$\min_{\boldsymbol{w}} L(\boldsymbol{w}, \boldsymbol{a}) = \min_{\boldsymbol{w}} \quad \frac{1}{2}||\boldsymbol{w}||^2 - \sum_n a_n[y_n\, \boldsymbol{w}^T \phi(\boldsymbol{x_n}) - 1]$$

- Set derivative to 0:

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \implies \boldsymbol{w} = \sum_n a_n y_n \phi(\boldsymbol{x}_n)$$

- Substitute $\boldsymbol{w}$ by $\sum_n a_n y_n \phi(\boldsymbol{x}_n)$ in $L(\boldsymbol{w}, \boldsymbol{a})$ to obtain:

$$L(\boldsymbol{a}) = \sum_n a_n - \frac{1}{2}\sum_n \sum_{n'} a_n a_{n'} y_n y_{n'} k(\boldsymbol{x}_n, \boldsymbol{x}_{n'})$$

Dual Problem

- We are then left with an optimization in $\boldsymbol{a}$ only known as the **dual problem**

$$\max_{\boldsymbol{a}} L(\boldsymbol{a})$$

$$\text{s.t. } a_n \geq 0$$

- **Sparse optimization**: many a_n's are 0

Classification

- Primal problem:

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

- Dual problem:

$$y_* = sign\left(\sum_n a_n y_n \phi(\boldsymbol{x}_n)^T \phi(\boldsymbol{x}_*)\right)$$

$$y_* = sign\left(\sum_n a_n y_n k(\boldsymbol{x}_n, \boldsymbol{x}_*)\right)$$

Generalization

- Support vector machines generalize quite well
  - i.e., overfitting is rare

- Reason: maximizing the margin is equivalent to minimizing an upper bound on the worst-case loss (worst loss for any underlying input distribution).

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 $\xi_n \geq 0$

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

- Picture:

Soft margin classifier

- New optimization problem:

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

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

$$\text{and} \quad \xi_n \geq 0 \qquad \forall n$$

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

UNIVERSITY OF WATERLOO

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 $C \rightarrow \infty$, 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 Classification

- Optimization problem:

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

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

- Equivalent to binary SVM when we have only two classes

Overlapping classes

- Add slack variables:

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

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

- Equivalent to binary SVM when we have only two classes

Public Lecture

- Speaker: Pascal Poupart
- Title: **From AlphaGo to ChatGPT**
- Date: April 12 @ 1:30 pm
- Location: DC1350

Other AI Courses

- CS486/686: Intro to AI (S23 instructor: Pascal Poupart)
  - includes reinforcement learning, causality, decision making
- CS485/685: Learning theory
- CS484/684: Computer vision
- CS479: Biologically plausible neural networks
- CS794: Optimization for Data Science
- CS885: Reinforcement Learning (instructor: Pascal Poupart)
- CS886: Advanced topics in AI
  - Graph neural networks, NLP, Vision, multiagent systems, robust ML, learning theory