CS540 Introduction to Artificial Intelligence

Lecture 5

Young Wu

Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

June 6, 2022

Guess the Percentage

Admin

The Percentage

Admin

Guess the Percentage

Admin

Sharing Solutions

Admin

Shared Solution List and Feedback

Admin

Maximum Margin Diagram

Motivation

SVM Weights

Quiz

SVM Weights Diagram

Quiz

SVM Weights

Quiz

SVM Weights Diagram

Quiz

Constrained Optimization Diagram

Definition

Constrained Optimization Derivation

Definition

Soft Margin Diagram

Definition

Soft Margin Derivation

Definition

SVM Formulations

Definition

Soft Margin

Quiz

Soft Margin 2

Quiz

Subgradient Descent

Definition

$$\min_{w} \frac{\lambda}{2} w^T w + \frac{1}{n} \sum_{i=1}^{n} \max\left\{0, 1 - (2y_i - 1)\left(w^T x_i + b\right)\right\}$$

- The gradient for the above expression is not defined at points with $1 - (2y_i - 1)\left(w^T x_i + b\right) = 0$.
- Subgradient can be used instead of a gradient.

Subgradient 1

Quiz

Subgradient 2

Quiz

Subgradient Descent Step

Definition

- One possible set of subgradients with respect to w and b are the following.

$$\partial_w C \ni \lambda w - \sum_{i=1}^{n} (2y_i - 1)\, x_i \mathbb{1}_{\{(2y_i - 1)(w^T x_i + b) \geqslant 1\}}$$

$$\partial_b C \ni - \sum_{i=1}^{n} (2y_i - 1)) \mathbb{1}_{\{(2y_i - 1)(w^T x_i + b) \geqslant 1\}}$$

- The gradient descent step is the same as usual, using one of the subgradients in place of the gradient.

Regularization Parameter

Definition

$$w = w - \alpha \sum_{i=1}^{n} z_i \mathbb{1}_{\{z_i w^T x_i \geqslant 1\}} x_i - \lambda w$$

$$z_i = 2y_i - 1, i = 1, 2, ..., n$$

- λ is usually called the regularization parameter because it reduces the magnitude of w the same way as the parameter λ in $L2$ regularization.
- The stochastic subgradient descent algorithm for SVM is called PEGASOS: Primal Estimated sub-GrAdient SOlver for Svm.

Kernel Trick $1D$ Diagram

Motivation

Kernelized SVM

Definition

- With a feature map φ, the SVM can be trained on new data points $\{(\varphi(x_1), y_1), (\varphi(x_2), y_2), ..., (\varphi(x_n), y_n)\}$.
- The weights w correspond to the new features $\varphi(x_i)$.
- Therefore, test instances are transformed to have the same new features.

$$\hat{y}_i = \mathbb{1}_{\{w^T\varphi(x_i) \geqslant 0\}}$$

Kernel Trick for XOR

Quiz

Kernel Trick for XOR

Quiz

Kernel Matrix

Definition

- The feature map is usually represented by a $n \times n$ matrix K called the Gram matrix (or kernel matrix).

$$K_{ii'} = \varphi\left(x_i\right)^T \varphi\left(x_{i'}\right)$$

Examples of Kernel Matrix

Definition

- For example, if $\varphi\left(x\right) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$, then the kernel matrix can be simplified.
$$K_{ii'} = \left(x_i^T x_{i'}\right)^2$$
- Another example is the quadratic kernel $K_{ii'} = \left(x_i^T x_{i'} + 1\right)^2$. It can be factored to have the following feature representations.
$$\varphi\left(x\right) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1\right)$$

Examples of Kernel Matrix Derivation

Definition

Popular Kernels

Discussion

- Other popular kernels include the following.

1. Linear kernel: $K_{ii'} = x_i^T x_{i'}$
2. Polynomial kernel: $K_{ii'} = \left(x_i^T x_{i'} + 1\right)^d$
3. Radial Basis Function (Gaussian) kernel:
$$K_{ii'} = \exp\left(-\frac{1}{\sigma^2}\left(x_i - x_{i'}\right)^T\left(x_i - x_{i'}\right)\right)$$

- Gaussian kernel has infinite-dimensional feature representations. There are dual optimization techniques to find w and b for these kernels.

Kernel Matrix

Quiz

Kernel Matrix Math

Quiz

Kernel Matrix 2

Quiz

Kernel Matrix Math 2

Quiz