CS540 Introduction to Artificial Intelligence

Lecture 5

Young Wu

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

July 6, 2022

Adverse Selection
Quiz

- Suppose the last two digits of your 10-digit student ID is the expected grade (out of 100) you will get in a course. Choose between the two courses:
- A : a course in which you get your expected grade.
- B : a course in which you get the average expected grade of everyone taking this course.

Adverse Selection, ID

Quiz

- Enter the last two digits of your ID.

Course Rhythm

Admin

1. (Q) In-class Quizzes, 0.5 points $(T - F)$
2. (D) Group discussion (reply to the Discussion post, also make it resolved), 0.5 points (M)
3. (D) Sharing solutions (create a note, not question, and tag $m2, m3, d1)$, 0.5 points each (M)
4. (M) Math homework, 1 point (M)
5. (P) Programming homework, 8 points (M)
6. (X) Exams, see Midterm page for past exams (same format this year).

Course Grades

Admin

- Final grade $= 0.3 \cdot X + 0.1 \max(X, Q) + 0.1 \max(X, D) + 0.1 \max(X, M) + 0.4 \cdot P$
- Additional discussion points used in borderline grades (for example 89 to A).

Sharing Solutions

Admin

1. Use LaTeX (Word, Maple, MyScript etc).

   sqrt((a_1^2) / (2 pi)) is difficult to read compared to $\sqrt{\frac{a_1^2}{2\pi}}$.
2. Handwritten on tablet or on paper and photo or scan (Office Lens).
3. Other suggestions?

Sharing Solutions

- For solution sharing, please make sure it is Piazza note, not a Piazza question.
- For actual questions, please use a different name, *e.g.*"$M2Q1$ Question" or "Question about $M2Q1$".
- Make sure you tag the post correctly: $m2$, $m3$, or $d1$ in order to get the points.
- Please sign up before making the post and please do not sign up for more than 4 questions per week.
- I will either "good note" the post or leave a comment: if I leave a comment, please update your answers, reply to my comment, and remember to make the reply "unresolved" so I can see.

Maximum Margin Diagram

Motivation

SVM Weights

Quiz

- Find the weights w_1, w_2 for the SVM classifier $\mathbb{1}_{\{w_1 x_{i1} + w_2 x_{i2} + 1 \geqslant 0\}}$ given the training data $x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $x_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ with $y_1 = 1, y_2 = 0$.
- $A : w_1 = 0, w_2 = -2$
- $B : w_1 = -2, w_2 = 0$
- $C : w_1 = -1, w_2 = -1$
- $D : w_1 = -2, w_2 = -2$

SVM Weights Diagram

Quiz

SVM Weights

Quiz

- Find the weights w_1, w_2 for the SVM classifier $\mathbb{1}_{\{w_1 x_{i1} + w_2 x_{i2} + 1 \geqslant 0\}}$ given the training data $x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, x_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, x_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ with $y_1 = 1, y_2 = y_3 = 0$.
- $A : w_1 = -1.5, w_2 = -1.5$
- $B : w_1 = -2, w_2 = -1.5$
- $C : w_1 = -1.5, w_2 = -2$
- $D : w_1 = -2, w_2 = -2$
- $E : I$ don't understand SVM

SVM Weights Diagram

Quiz

Constrained Optimization Diagram

Definition

Constrained Optimization Derivation

Definition

- The goal is to maximize the margin subject to the constraint that the plus plane and the minus plane separates the instances with $y_i = 0$ and $y_i = 1$.

$$\max_{w} \frac{2}{\sqrt{w^T w}} \text{ such that } \begin{cases} \left(w^T x_i + b\right) \leqslant -1 & \text{if } y_i = 0 \\ \left(w^T x_i + b\right) \geqslant 1 & \text{if } y_i = 1 \end{cases}, i = 1, 2, ..., n$$

- This is equivalent to the following minimization problem, called hard margin SVM.

$$\min_{w} \frac{1}{2} w^T w \text{ such that } (2y_i - 1)\left(w^T x_i + b\right) \geqslant 1, i = 1, 2, ..., n$$

Soft Margin Diagram

Definition

Soft Margin Derivation

Definition

SVM Formulations

Definition

- Hard margin:

$$\min_{w} \frac{1}{2} w^T w \text{ such that } (2y_i - 1)\left(w^T x_i + b\right) \geqslant 1, i = 1, 2, ..., n$$

- Soft margin:

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

Soft Margin

Quiz

- Let $w = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $b = 3$. For the point $x = \begin{bmatrix} 4 \\ 5 \end{bmatrix}, y = 0$, what is the smallest slack variable ξ for it to satisfy the margin constraint?

$$(2y_i - 1)\left(w^T x_i + b\right) \geqslant 1 - \xi_i, \xi_i \geqslant 0$$

Soft Margin 2

Quiz

- Let $w = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $b = 3$. For the point $x = \begin{bmatrix} -4 \\ -5 \end{bmatrix}$, $y = 0$, what is the smallest slack variable ξ for it to satisfy the margin constraint?
- A : -10
- B : 0
- C : 10
- D : None of the above
- E : I don't understand what is ξ

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

- Which ones are subderivatives of $\max \{x, 0\}$ at $x = 0$?
- $A : -1$
- $B : -0.5$
- $C : 0$
- $D : 0.5$
- $E : 1$

Subgradient 2

Quiz

- Which ones are subderivatives of $\left|x\right|$ at $x = 0$?
- $A : -1$
- $B : -0.5$
- $C : 0$
- $D : 0.5$
- $E : 1$

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

- SVM with quadratic kernel $\varphi\left(x\right) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$ can correctly classify the following training set?

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

Kernel Trick for XOR

Quiz

- SVM with kernel $\varphi(x) = (x_1, x_1 x_2, x_2)$ can correctly classify the following training set.

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

- A : True.
- B : False.

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 $\varphi(x) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$, 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.

$$\varphi(x) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1\right)$$

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

- What is the feature vector $\varphi(x)$ induced by the kernel $K_{ii'} = \exp(x_i + x_{i'}) + \sqrt{x_i x_{i'}} + 3$?

Kernel Matrix Math

Quiz

Kernel Matrix 2

Quiz

- What is the feature vector $\varphi(x)$ induced by the kernel $K_{ii'} = 4\exp(x_i + x_{i'}) + 2x_i x_{i'}$?
- A : $(4\exp(x), 2\sqrt{x})$
- B : $\left(2\exp(x), \sqrt{2}\sqrt{x}\right)$
- C : $(4\exp(x), 2x)$
- D : $\left(2\exp(x), \sqrt{2}x\right)$
- E : None of the above

Kernel Matrix Math 2

Quiz