Recitation 2: kNN, Perceptron & Linear Regression

10-301/10-601 Introduction to Machine Learning (Summer 2024)
http://www.cs.cmu.edu/~hchai2/courses/10601

1 kNN & Perceptron

1.1 *k* NN

1. Using the figure below, what would you categorize the green circle as with $k = 3$? $k = 5$?

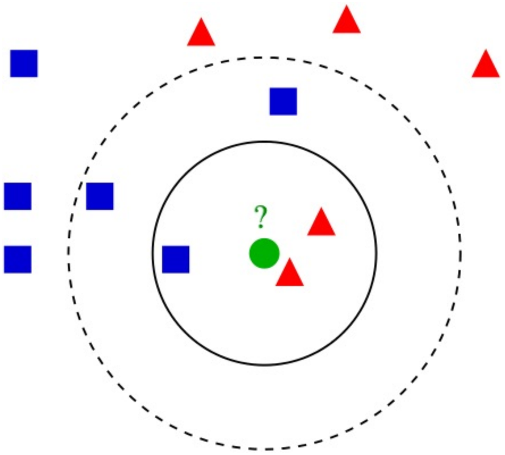
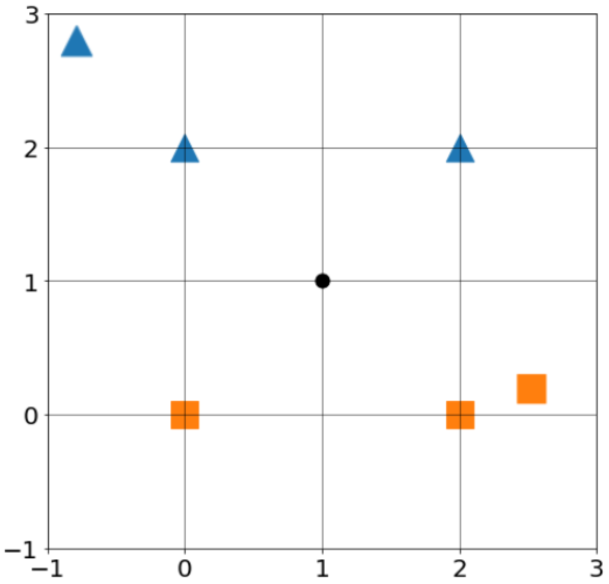


Figure 1: An example of k-NN on a small dataset; image source from Wikipedia

Your answer:

2. **Select all that apply:** Consider a binary k-NN classifier where $k = 4$ and the two labels are "triangle" and "square". Consider classifying a new point $\mathbf{x} = (1, 1)$, where two of the $\mathbf{x}$'s nearest neighbors are labeled "triangle" and two are labeled "square" as shown below.



Which of the following methods can be used to break ties or avoid ties on this dataset?

□ Assign $\mathbf{x}$ the label of its nearest neighbor

□ Flip a coin to randomly assign a label to $\mathbf{x}$ (from the labels of its 4 closest points)

□ Use $k = 3$ instead

□ Use $k = 5$ instead

□ None of the above.

3. Consider the following data concerning the relationship between academic performance and salary after graduation. High school GPA and university GPA are two numerical features and salary is the numerical target. Note that salary is measured in thousands of dollars per year.

Student ID	**High School GPA**	**University GPA**	**Salary**
1	2.5	3.8	45
2	3.3	3.5	90
3	4.0	4.0	142
4	3.0	2.0	163
5	3.8	3.0	2600
6	3.3	2.8	67
7	3.9	3.8	unknown

(a) Among Students 1 to 6, who is the nearest neighbor to Student 7, using Euclidean distance?

Nearest Neighbor

Work

(b) Now, our task is to predict the salary Student 7 earns after graduation: using $k = 3$, what is the average salary of Student 7's k nearest neighbors, rounded to the nearest integer?

Salary

Work

(c) **Select all that apply:** Suppose that the first 6 students shown above are only a subset of your full training data set, which consists of 10,000 students. We apply kNN using Euclidean distance to this problem and we define the loss function on this full data set to be the mean squared error (MSE) of salary. Now consider the possible consequences of modifying the data in various ways. Which of the following changes **could** have an effect on training loss over the full data set as measured by mean squared error (MSE) of salary?

□ Rescaling only “High School GPA” to be a percentage of 4.0

□ Rescaling only “University GPA” to be a percentage of 4.0

□ Rescaling both “High School GPA” and “University GPA”, so that each is a percentage of 4.0

□ None of the above.

1.2 Perceptron Mistake Bound

If a dataset has margin γ and all points inside a ball of radius R, then the perceptron makes less than or equal to $(R/\gamma)^2$ mistakes.

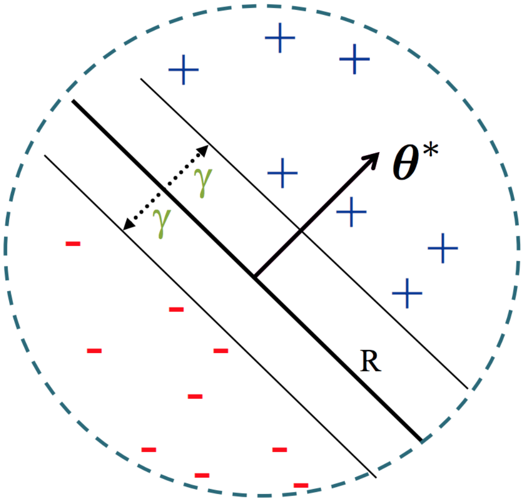


Figure 2: Perceptron Mistake Bound Setup

Definitions:

- Margin:
  - The margin of example x wrt a linear separator w is the (absolute) distance from x to the plane $w \cdot x = 0$.
  - The margin γ_w of a set of examples S wrt a linear separator w is the smallest margin over points $x \in S$.
  - The margin γ of a set of examples S is the maximum γ_w over all linear separators w.
- Linear Separability: For a binary classification problem, a set of examples S is linearly separable if there exists a linear decision boundary that can separate the points.
- Update Rule: When the k-th mistake is made on data point $\mathbf{x}^{(i)}$, the parameter update is

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \mathbf{y}^{(i)}\mathbf{x}^{(i)}$$

- Radius: The radius is the maximum distance between a point and the origin in a dataset, defining the radius of a circle centered at the origin that encompasses all points of the dataset.

We say the (batch) perceptron algorithm has *converged* when it stops making mistakes on the training data.

1. **Main Takeaway:** What does the Perceptron Mistake Bound imply about linearly separable training datasets?

Your answer:

2. **Select all that apply:** Which of the following is/are correct statement(s) about the mistake bound of the perceptron algorithm?

   □ If the minimum distance from any data point to the separating hyperplane is increased, without any other change to the data points, the mistake bound will also increase.

   □ If the whole dataset is shifted away from origin, then the mistake bound will also increase.

   □ If the pair-wise distance between data points is increased, i.e. the data is scaled by some constant value, then the mistake bound will also increase.

   □ The mistake bound is linearly inverse-proportional to the minimum distance of any data point to the separating hyperplane of the data.

   □ None of the above.

3. The following problem will walk you through an application of the Perceptron Mistake Bound. The following table shows a linearly separable dataset, and your task will be to determine the mistake bound for the dataset.

   NOTE: The proof of the perceptron mistake bound requires that the optimal linear separator passes through the origin. To make the linear separator pass through the origin, we fold the bias into the weights and prepend a 1 to each training example's input. The original data is on the left, and the result of this prepending is shown on the right. **Be sure to use the modified dataset on the right in your calculations.**

x_1	x_2	y
-2	2	1
-1	-3	-1
-2	-3	-1
0	1	1
2	-1	1

$\boldsymbol{x_0}$	x_1	x_2	y
1	-2	2	1
1	-1	-3	-1
1	-2	-3	-1
1	0	1	1
1	2	-1	1

(a) Compute the radius R of the "circle" centered at the origin that bounds the data points.

Radius:

Work

(b) Assume that the linear separator with the largest margin is given by

$$\boldsymbol{\theta}^{*T} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} = 0, , \text{ where } \boldsymbol{\theta}^* = \begin{bmatrix} 6 \\ 3 \\ 4 \end{bmatrix}$$

Now, compute the margin of the dataset.

Margin:

Work

(c) Based on the above values, what is the theoretical perceptron mistake bound for this dataset, given this linear separator?

Mistake Bound:	Work

2 Linear Regression

2.1 Objective Functions

1. In the context of linear regression, what does an objective function $\ell(\mathbf{w})$ do?

Your answer:

2. What are some desirable properties of a good objective function?

Your answer:

2.2 Closed-form Solution for Linear Regression

Suppose we are given the following dataset where x is the input and y is the output:

x	1.0	2.0	3.0	4.0	5.0
y	2.0	4.0	7.0	8.0	11.0

Based on our inductive bias, we think that the linear hypothesis with no intercept should be used here. We also want to use the Mean Squared Error as our objective function: $\frac{1}{5}\sum_{i=1}^{5}(y^{(i)} - wx^{(i)})^2$, where $y^{(i)}$ is our i^{th} data point and w is our weight. Using the closed-form method, find w.

1. What is the closed-form formula for w?

Your answer:

2. What is the value of w?

Your answer:

Now let’s extend the data set to include more features, $\mathbf{x} \in \mathbb{R}$:

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$
x_1	1.0	2.0	3.0	4.0	5.0
x_2	-2.0	-5.0	-6.0	-8.0	-11.0
x_3	3.0	8.0	9.0	12.0	14.0
y	2.0	4.0	7.0	8.0	11.0

We again think that a linear hypothesis with no bias should be used here. We also want to use the Mean Squared Error as our objective function:

$$\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2,$$

where $\mathbf{w} = [w_1, w_2, w_3]^T$, $\mathbf{x}^{(i)}$ is the i^{th} datapoint and $y^{(i)}$ is the i^{th} $y-$value.

1. What are the design matrix X and target vector $\mathbf{y}$ in this setting?

Your answer:

2. What is the closed-form matrix solution for $\mathbf{w}$?

Your answer: