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 $k = 3$? $k = 5$?

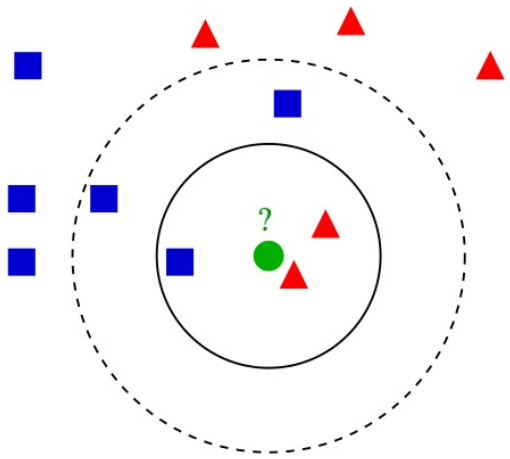
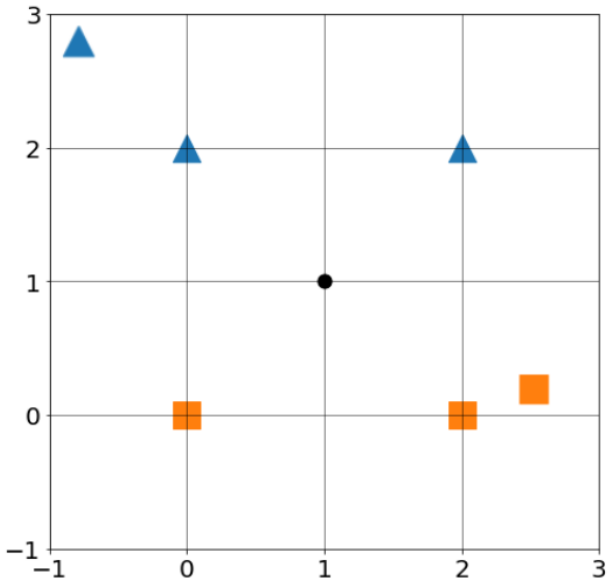


Figure 1: An example of k-NN on a small dataset; image source from Wikipedia

Your answer:

Example of k-NN classification. The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles.
If k = 3 (solid line circle) it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle.
If k = 5 (dashed line circle) it is assigned to the first class (3 squares vs. 2 triangles inside the outer circle).

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.

B, D

1. is false because points at (0,0), (2,0), (0,2), and (2,2) all have the same distance to x = (1,1).
2. is true because it would break the tie (this strategy is introduced in class).
3. is false because there are four points that have the same distance to x, and we don't have a strategy to pick which three to choose.
4. is true because with k = 5, we will take the square at (2.7,0.6) into consideration, which will lead to x being labelled square (3/5).

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

3

Work

Distance from student i to student 7 is listed below, $i \in \{1, .., 6\}$. We can compute the Euclidean distance by applying $d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$. For example, d(student 1, student 7) = $\sqrt{(2.5 - 3.9)^2 + (3.8 - 3.8)^2} = 1.40$.

- Student 1: distance = 1.40
- Student 2: distance = 0.67
- Student 3: distance = 0.22
- Student 4: distance = 2.01
- Student 5: distance = 0.81
- Student 6: distance = 1.17

As we can see, student 3 has the shortest distance to student 7, and that makes it student 7's nearest neighbor

(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

944

Work

From the list computed in the solution to previous problem, student 7's three nearest neighbors are: students 2, 3, and 5. There corresponding incomes are: 90, 142, and 2600. Applying the knn regression algorithm, our prediction for student 9's salary is $\frac{90+142+2600}{3} = 944$.

(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.

A and B.

1. True, by only scaling one feature (High School GPA), it will distort the order of original distances, which will affect our classification and training loss. (Imagine this extreme case: multiplying High School GPA by 0.00001, this will basically make this feature negligible, and the distance will solely on University GPA.)
2. True, for the same reason as (a), except change the High School GPA to University GPA
3. False, if we scale both features by a percentage p. If we denote d' to be the scaled distance and d to be the original distance.

$$\begin{aligned} d'(x,y) &= \sqrt{(px_1 - py_1)^2 + (px_2 - py_2)^2} \\ &= \sqrt{p^2[(x_1 - y_1)^2 + (x_2 - y_2)^2]} \\ &= p\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \\ &= p * d(x,y) \end{aligned}$$

Since we are essentially multiple all distances by p, it won't change the order, thus not affecting the training loss.

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.

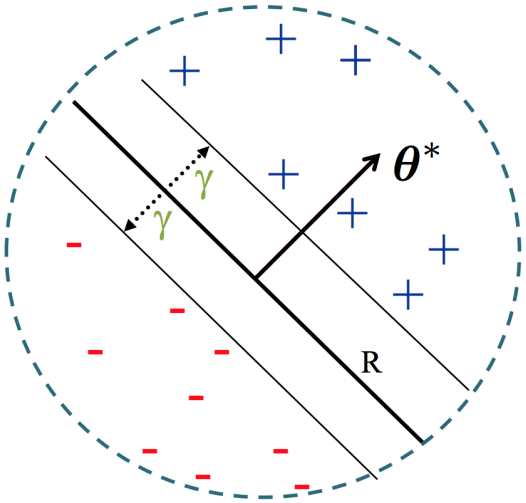


Figure 2: Perceptron Mistake Bound Setup

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:

   For linearly separable data, if the perceptron algorithm repeatedly cycles through the data, it will converge in a finite number of steps.

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.

   B. The perceptron mistake bound is given by $(R/\gamma)^2$, where R is radius of sphere that contains all points in the dataset. That is, R is the maximum distance from any point to the origin. Therefore we see that B is true.

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
3.7417	The radius is the distance of the furthest point from the origin. This is simply the point with the coordinates having the largest L2 norm. In this case, the point is (1, -2, -3), with magnitude $\sqrt{14} = 3.7417$.

   (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
1.0243	Apply the distance-from-a-point-to-a-line formula for each point in the dataset. The smallest value is the margin. (1, -2, 2) and (1, 2, -1) are equally close, with distance to the linear separator: $d = \frac{6 + 3 * -2 + 4 * 2}{\sqrt{6^2 + 3^2 + 4^2}} \approx 1.0243$

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

Mistake Bound:
13.3439

Work

This problem is plug and chug. The Block, Novikoff Perceptron Mistake Bound formula covered in lecture puts the bound at $(R/\gamma)^2$. Using the previous two answers, this computation is trivial - accept answers that are slightly off due to rounding

$$MistakeBound = (3.7417/1.0243)^2 \approx 13.3439$$

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:

Measures how "bad" a particular linear model is.

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

Your answer:

- Should be differentiable
- Preferably convex

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:

$$\ell(w) = \frac{1}{5}\sum_{i=1}^{5}(y^{(i)} - wx^{(i)})^2 \rightarrow \frac{d\ell(w)}{dw} = \frac{1}{5}\sum_{i=1}^{5} -2x^{(i)}(y^{(i)} - wx^{(i)}) = 0$$

$$\sum_{i=1}^{5} x^{(i)}(y^{(i)} - wx^{(i)}) = 0$$

$$\sum_{i=1}^{5} x^{(i)}y^{(i)} - \sum_{i=1}^{5} w(x^{(i)})^2 = 0$$

$$w\sum_{i=1}^{5}(x^{(i)})^2 = \sum_{i=1}^{5} x^{(i)}y^{(i)} \rightarrow w = \frac{\sum_{i=1}^{5} x^{(i)}y^{(i)}}{\sum_{i=1}^{5}(x^{(i)})^2}$$

2. What is the value of w?

Your answer:

$$\sum_{i=1}^{5} x^{(i)}y^{(i)} = 118$$

$$\sum_{i=1}^{5}(x^{(i)})^2 = 55$$

$$w = \frac{\sum_{i=1}^{5} x^{(i)}y^{(i)}}{\sum_{i=1}^{5}(x^{(i)})^2} = \frac{118}{55} = 2.15$$

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:

The design matrix X is given by:

$$\begin{bmatrix} 1 & -2 & 3 \\ 2 & -5 & 8 \\ 3 & -6 & 9 \\ 4 & -8 & 12 \\ 5 & -11 & 14 \end{bmatrix}$$

and the target vector $\mathbf{y}$ is given by:

$$\begin{bmatrix} 2, 4, 7, 8, 11 \end{bmatrix}^T$$

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

Your answer:

Using the closed-form formula in class

$$\hat{\mathbf{w}} = (X^T X)^{-1} X^T Y$$

we get

$$\mathbf{w} = \begin{bmatrix} 2.36 \\ -0.205 \\ -0.218 \end{bmatrix}$$