CS 129 – Winter 2026

Problem Set 3

Instructors: Andrew Ng, Younes Bensouda Mourri

Due: 11:59pm Tuesday March 10th, 2026

Problem 1: Decision Trees

Consider this small dataset. We will be using decision trees in this problem to predict whether or not a suggested dish will make it onto the menu in our restaurant.

Dish name	Prep Time	Num Ingredients	Num of cooks who liked it	Prep Cost	On the Menu
Crispy Chicken	20	6	10	5	Yes
Loaded Salad	10	17	15	3	No
Triple Burger	15	8	8	7	Yes
Quad Burger	15	9	3	9	No
Banana split	2	10	18	4	Yes

1. (2 points) How deep does a decision tree like the one we saw in class and on Coursera need to be to perform as a perfect classifier on this dataset?

   Your Solution Here

2. (3 points) Draw such a tree. (There is more than one right answer)

   Your Solution Here

3. (4 points) Does normalizing your inputs help with decision trees in the same way it would help with other algorithms? Could it hurt the performance?

   Your Solution Here

4. (3 points) In this assignment we introduce the idea of decision trees with multiple branches. This can be useful for categorical features where each category splits into a branch or can be used to set $n-1$ thresholds for n branches for a continuous value.
   The measurement of information gain for such a tree is calculated by

$$H(p_1^{root}) - \sum_{i}^{n} w^i H(p_1^i)$$

   instead of

$$H(p_1^{root}) - w^{right} H(p_1^{right}) + w^{left} H(p_1^{left})$$

   Draw a tree with three branches trained on our dataset to be a perfect classifier. (There is only one possible tree this time.)

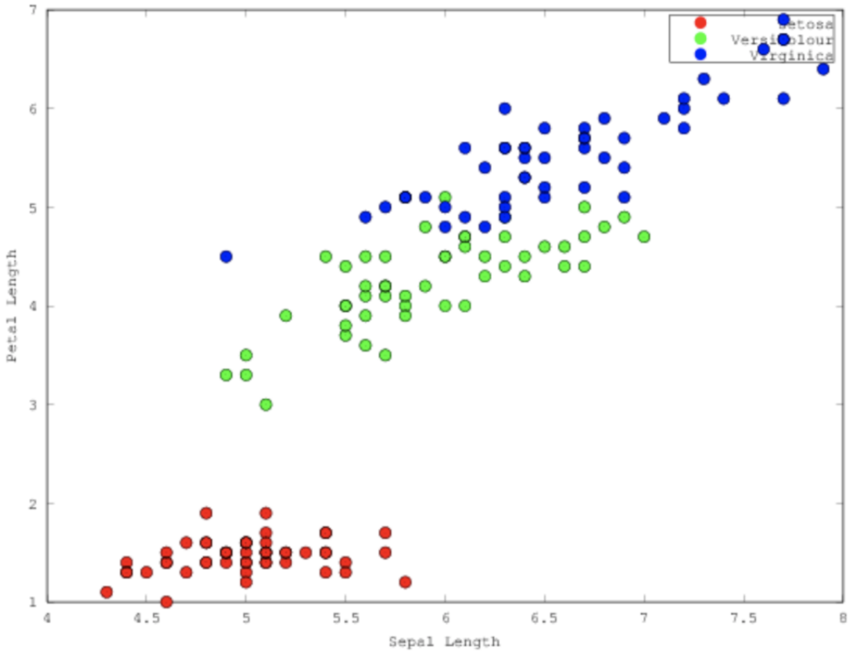
   Your Solution Here

5. (2 points) Now draw a decision tree with two branches that uses the same decision logic and splits as the three branched one. (There are two right answers)

   Your Solution Here

As you may have observed from the previous examples adding more branches and adding more depth are practically equivalent, and both can lead to over-fitting.

6. This is a graph with a feature on the x axis and another feature on the y axis.



(3 points) What would the decision boundary look like for a decision tree with a maximum depth of two trained on recognising Versicolour?

Your Solution Here

7. (3 points) Generally describe what the decision boundary would look like if we used a maximum depth of 20.

Your Solution Here

Problem 2: Principal Component Analysis

1. Choose **True/False**. No justification needed.

   (a) (2 points) The goal of PCA is to interpret the underlying structure of the data in terms of the principal components that are best at predicting the output variable.
   Your Solution Here

   (b) (2 points) The sum of the PCA eigenvalues is equal to the sum of the variances of the variables. (Hint: think about the trace of a covariance matrix.)
   Your Solution Here

   (c) (2 points) Principal component analysis (PCA) can be used with variables of any mathematical types: quantitative, qualitative, or a mixture of these types.
   Your Solution Here

2. A classmate says: *"PCA reduces the dimensionality of our dataset by identifying and keeping the most important subset of the original features."* Which of the following responses is most accurate?

   - A: The classmate is correct as PCA ranks the original features by variance and discards the least important ones.
   - B: The classmate is incorrect as PCA reduces dimensionality by constructing a new set of orthogonal axes that are linear combinations of *all* original features, ordered by the variance they explain.
   - C: The classmate is correct as PCA selects a subset of original features, but uses covariance rather than variance to rank them.
   - D: The classmate is incorrect as PCA reduces dimensionality by randomly projecting the data onto a lower-dimensional subspace to preserve pairwise distances.

   Your Solution Here

3. (2 points) Give one advantage and one disadvantage of using PCA.
   Your Solution Here

4. (3 points) How can PCA be used to speed up supervised learning?
   Your Solution Here

Problem 3: K-means algorithm

Suppose we have the following points in one dimension:

$$x_1 = 0,\ x_2 = 2,\ x_3 = 3,\ x_4 = 8,\ x_5 = 10$$

Run the 2-means clustering until convergence with the following initialization:

$$\mu_1 = -1,\ \mu_2 = 5$$

Note: in the case of a tie, assign the point to the class with a lower number (i.e. if one point is tied between class 1 and class 2, assign it to class 1).

1. (2 points) Draw a number line to help you visualize what is happening.

   Your Solution Here

2. (2 points) How many iterations did you perform? *Note*: do not double count! Therefore if iterations n and $n+1$ give the same result, the algorithm converges in n iterations.

   Your Solution Here

3. (5 points) What is the final assignment?

   Your Solution Here

4. (5 points) What are the final centroids?

   Your Solution Here

5. (5 points) Remember that the loss in the K-means algorithm is given by:

   $$\text{Loss} = \sum_{i=1}^{m} ||x_i - \mu_{z_i}||^2 \text{ with } z_i \text{ the cluster of point } i$$

   Compute the final loss.

   Your Solution Here

6. (1 point) In the general case, does the K-means algorithm converge to the global minimum?

7. (3 points) One challenge in K-means is choosing the number of clusters K. Describe the elbow method for selecting K. What are its limitations?

   Your Solution Here

8. (4 points) Another challenge is the sensitivity of K-means to initialization.

   (a) (2 points) Give an example or description of a case where random initialization leads to a poor final clustering.

   Your Solution Here

   (b) (2 points) K-means++ addresses this by choosing initial centroids with a probability proportional to their squared distance from the nearest already-chosen centroid. Why does this lead to better initializations than choosing centroids uniformly at random?

   Your Solution Here

Problem 4: Evaluating Clustering

Unlike supervised learning, clustering has no ground truth labels to evaluate against. In this problem we explore two common ways to measure clustering quality.

1. (3 points) The **inertia** (also called within-cluster sum of squares) is defined as the total loss from the K-means objective. What are the limitations of using inertia alone to evaluate clustering quality? In particular, what happens to inertia as K increases toward m (the number of data points)?

   Your Solution Here

2. (4 points) The **silhouette score** for a single point $x^{(i)}$ is defined as:

   $$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

   where $a(i)$ is the mean distance from $x^{(i)}$ to all other points in its cluster, and $b(i)$ is the mean distance from $x^{(i)}$ to all points in the nearest neighboring cluster.

   (a) (2 points) What is the range of $s(i)$? What do values close to 1, close to 0, and close to -1 indicate?

   Your Solution Here

   (b) (2 points) How could you use the silhouette score to choose K?

   Your Solution Here

3. (2 points) Suppose you run K-means with $K = 3$ on a dataset that contains three clusters: two tight, well-separated clusters of 10 points each, and one large, spread-out cluster of 200 points. You then rerun K-means with $K = 6$, splitting each of the original three clusters into two sub-clusters. Explain why the inertia for $K = 6$ will be much lower than for $K = 3$, even though $K = 3$ is arguably the more meaningful clustering. Why would the silhouette score be more informative here?

   Your Solution Here

Problem 5: Pac-Man

Consider a Pac-Man agent navigating a small grid. At each step, Pac-Man can move Up, Down, Left, or Right. It receives a reward of $+10$ for eating a pellet, -20 for being eaten by a ghost, and 0 otherwise.

1. (3 points) Identify the **state**, **action**, and **reward** in this Pac-Man setting. Give one concrete example of each.

   Your Solution Here

2. (2 points) **True/False.** No justification needed.

   (a) A Pac-Man agent with $\gamma = 0$ will prioritize eating the nearest pellet over avoiding a ghost two steps away.

   Your Solution Here

   (b) If Pac-Man always moves toward the nearest pellet (never considering ghosts), this is a valid policy.

   Your Solution Here

3. (2 points) Pac-Man is hovering near a ghost but there is a pellet one step away. Explain in one or two sentences why the exploration vs. exploitation tradeoff matters in this situation.

   Your Solution Here