Lecture 11

Classification

Classification is the process of predicting the class label of given data points using a trained model. The classes are usually discrete (e.g., types of animals, spam vs. non-spam).

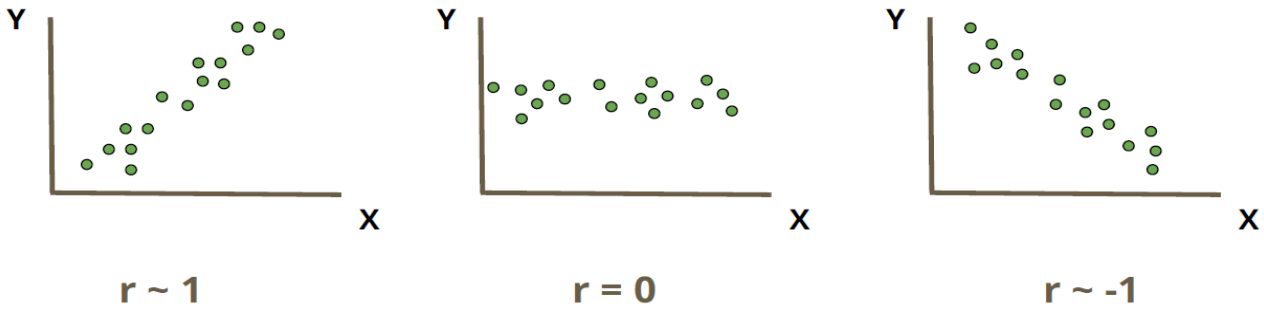
- There could be many correct answers
- There could be no correct answers
    - And maybe that's ok - no relationship is interesting information too
    - But the model could still be useful if it's correct most of the time
- Whether a task is feasible depends on:
    - The relationship between the predictors and the class

Correlation quantifies how strongly two variables are related and whether an increase (or decrease) in one variable tends to be associated with an increase or decrease in the other.

- Positive correlation: Variables increase or decrease together.
- Negative correlation: One variable increases while the other decreases.
- No correlation: Variables have no predictable relationship.

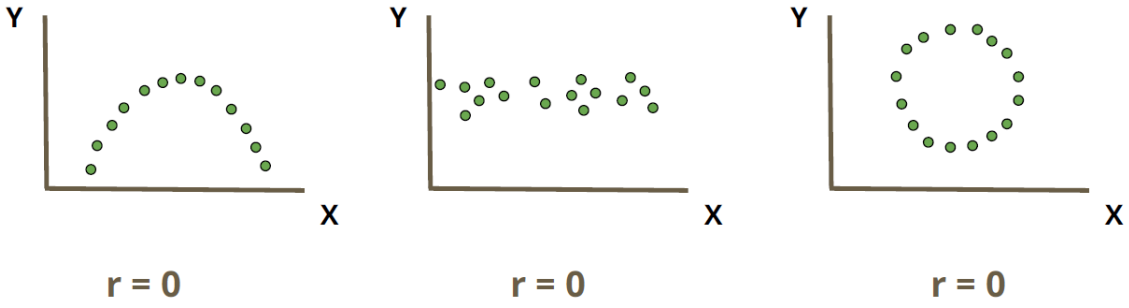
Pearson Correlation Coefficient (r): $r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$

Remember: Correlation does NOT mean causation.



- Left: $r \approx 1$: Strong positive correlation.
  As X increases, Y increases. Suggests X is a good linear predictor of Y.
- Middle: $r = 0$: No correlation.
  As X changes, Y does not show a pattern. X is not a good linear predictor of Y.
- Right: $r \approx -1$: Strong negative correlation.
  As X increases, Y decreases. X is a good linear predictor, but in the negative direction.

Good predictors often have high (positive or negative) correlation. If r is close to 0, the variable is probably not useful as a linear predictor.



No Correlation Doesn't Mean No Relationship

Even if $r = 0$, there might still be a relationship between X and Y, but not a linear one.

- Left: Quadratic relationship (but overall linear $r = 0$).
- Center: No relationship at all (random scatter, $r = 0$).
- Right: Circular pattern ($r = 0$).

Zero correlation only detects **lack of linear relationships**. Nonlinear relationships can exist even if $r = 0$, and may still be useful for prediction with the right tools.

Now usually, correlation measures how two numerical variables move together. But if Y is a class (categorical), we can't use the usual correlation coefficient.

Two Types of Categorical Data

1. *Nominal:* Categories without any order.
   Examples:
   a. Y = {Yes, No}
   b. Y = {Red, Green, Blue}
   c. Y = {London, Paris, NY}
   Key point: The categories are just names and have no natural order.

2. *Ordinal:* Categories with a natural order, but distances aren't meaningful.
   Example: Y = {Terrible, Bad, OK, Good, Great}
   Key point: There is a ranking, but the gap between each is unknown. For example, difference between Terrible and Bad may not be the same as between Good and Great.

So how do we measure if X (numeric) is a good predictor for categorical Y?

1. **For Nominal Y (no order):**
    - Instead of correlation, look at the mean/average of X for each category of Y.
    - For example: If Y is {Yes, No} and X is income, compare the average income for Yes and for No.
    - If these means are very different, X is a good predictor of Y. This is often done with boxplots, group means, or statistical tests like ANOVA.

2. **For Ordinal Y (ordered):**
    - You can assign numbers to each category (Terrible = 0, Bad = 1, OK = 2, etc.), which allows you to look for trends as X increases.
    - But you should not assume the distances are meaningful. There's order, but not true scale
    - However, for ordinal classes, it's often **better to compare their rank** instead of assigned values.
    - The actual gap between "Bad" and "OK" doesn't matter, only the order does.
    - Use statistical methods that compare positions/ranks (e.g. Spearman correlation).

Spearman Correlation Formula: $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$

Example:

Index	X	Y
1	10	1
2	20	0
3	30	2
4	40	3
5	50	4

Solution:

Rank means assigning a position to each value when sorted in ascending order.

For X: *X Values in ascending order:* 10, 20, 30, 40, 50 therefore, assign rank 1 to smallest (10), rank 5 to largest (50)

For Y: *Sort Y in ascending order:* 0, 1, 2, 3, 4 therefore, Assign rank 1 to smallest (0), rank 5 to largest (1)

Index	X	R(X)	Y	R(Y)
1	10	1	1	2
2	20	2	0	1
3	30	3	2	3
4	40	4	3	4
5	50	5	4	5

Now $d = R(X) - R(Y)$

$\boldsymbol{R(X)}$	$\boldsymbol{R(Y)}$	$\boldsymbol{d_i}$	$\boldsymbol{d_i^2}$
1	2	-1	1
2	1	1	1
3	3	0	0
4	4	0	0
5	5	0	0

Applying Spearman Formula: $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$

$$\rho = 1 - \frac{6(1+1+0+0+0)}{5(5^2-1)}$$

$$\rho = 1 - \frac{6(2)}{5(24)}$$

$$\rho = 1 - \frac{1}{10} = \frac{9}{10} = \mathbf{0.9}$$

Correlation vs Causation

1. *Temperature and Ice Cream Sales*

   Observation: Temperature and ice cream sales tend to increase together (positive correlation).

   Interpretation:

   - When temperature rises, ice cream sales spike (people buy more ice cream when it's hot).
   - However, this doesn't mean that ice cream sales cause temperature to increase—causation is one-way.

   Caveat: In places like deserts without ice cream availability, temperature still rises, but ice cream sales do not spike because there simply is no ice cream sold.

   Summary:

   - Temperature increase → causes → Ice cream sales increase.
   - Ice cream sales increase → does NOT cause → temperature increase.

2. *Sleeping with Shoes on and Headaches*

   Observation: People who sleep with shoes on tend to wake up with headaches (strong correlation).

   Key Insight:

   - Sleeping with shoes on does not cause headaches.
   - Waking up with headaches does not cause people to sleep with their shoes on.

   Explanation:

   - There is a third factor causing both: for example, going to bed drunk.
   - Being drunk causes people to sleep with shoes on AND to wake up with a headache.

   Summary:

   - Correlation exists, but no direct causation between the two variables.
   - Both are caused by a hidden third variable.

Testing for causality requires specific testing / experimentation with a control group.

But it's very hard to show that things are causally linked through observational data, especially if the relationship isn't deterministic.

Distinguishing Correlation and Causation Using Counterfactual Reasoning

- Distinguish **correlation** (variables move together) from **causation** (one variable directly influences another).
- Causation inference requires comparing outcomes **with and without** exposure to a factor (i.e., exposed vs unexposed groups).
- Use **probabilistic reasoning** and **counterfactuals** (what would happen if exposure did not exist) to estimate the causal effect.
- This approach helps quantify how much of an outcome is attributable to the exposure beyond natural baseline risk.
- Establishing causation from observational data involves careful analysis beyond simple correlation measures.

Instance-Based Classifiers

Instance-Based Classifiers are a category of classification algorithms in machine learning that do not learn an explicit global model. Instead, they classify new instances based on comparisons to instances from the training data.

- They store the training data (instances) and make predictions by comparing new inputs to these stored examples.
- Sometimes called lazy learners because the generalization only happens at the time of prediction, not during training.

Advantages	Disadvantages
Simple to implement and understand.	Computational cost grows with training data size.
Naturally adapts to underlying data structure.	Performance depends heavily on choice of distance metric and k.
Works well in multi-class and multi-modal problems.	Sensitive to noisy data and irrelevant features.

Example:

- **k-Nearest Neighbours (k-NN):** Most popular instance-based classifier.
- **Locally Weighted Learning:** Weighted combination of neighbours based on distance.
- **Case-Based Reasoning (CBR):** Uses stored cases to solve new problems.

K Nearest Neighbour Classifier

- K-NN is a simple, intuitive, and widely used instance-based classification algorithm.
- It classifies a new data point based on the classes of its K closest neighbours in the training data.

Working:

1. **Choose k:** Decide on the number of neighbours k to consider (e.g., 3, 5, 7).
2. **Select a distance metric:** Common choices include:
    a. Euclidean distance
    b. Manhattan distance
    c. Minkowski distance
    d. Others, depending on data type.
3. **Find neighbours:** For a new test instance:
    a. Compute the distance between the test instance and every training instance.
    b. Select the k training instances with the smallest distances.
4. **Predict class:**
    a. Majority vote: Assign the class most common among the k neighbours.
    b. Weighted vote: Weight neighbours' votes by their distance (closer neighbours have more influence).
5. **Example:**
    a. Suppose $k = 3$, the 3 nearest neighbours to a test point belong to classes: {A, B, A}.
    b. Majority class is A, so the test point is classified as A.

Disadvantages:

- Computationally expensive during prediction (distance to all training samples).
- Performance depends heavily on:
    - Choice of k
    - Distance metric
    - Feature scaling (normalization matters)
- Sensitive to noisy, irrelevant features, and outliers.
- No explicit model or explanation.

Scaling

- Why: Different features can have very different ranges (e.g., Age: 0-100 vs Income: 10k–1 million).
- Problem: Distance calculations can be dominated by features with larger numeric ranges, leading to biased results.
- Solution: Features need to be scaled to a similar range to ensure fair contribution to distance measures. (**Normalize or standardize** features)

Min – Max Scaling:

Rescales data to a fixed range, usually [0, 1].

Formula for a value x : $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$

Example: $X = [20,30,40,50,60]$

$$\min(X) = 20 \, ; \max(X) = 60$$

For $x = 30$:

$$x' = \frac{30 - 20}{60 - 20} = \frac{10}{40} = 0.25$$

Similarly do for 40 and 50, and finally scaled would be:

$$X' = [0,0,25.25,0.5,0.75,1]$$

Standardization:

Centres data around zero mean and scales it to unit variance.

Formula for a value x: $z = \frac{x - \mu}{\sigma}$

where μ is the mean and σ is the standard deviation of X.

Example: $X = [20,30,40,50,60]$

Mean μ : $\frac{20+30+40+50+60}{5}$ = 40

Standard Deviation σ = 14.14

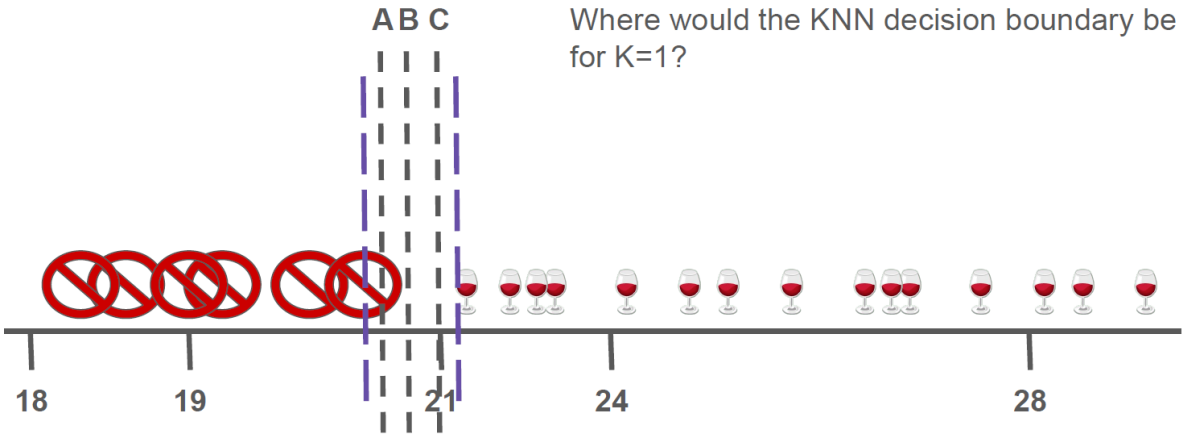
Therefore, $x = 30$:

$$z = \frac{30 - 40}{14.14} = -\frac{10}{14.14} = -0.707$$

Similarly do for 40 and 50, and finally scaled would be:

$$X' = [-1.41, -0.71, 0, 0.71, 1.41]$$

KNN can be problematic in high dimensions (curse of dimensionality)



Answer: C

- The last "No" (crossed circle) is at 20, and the first "Yes" (wine glass) is at 22.
- The decision boundary will be at the midpoint between 20 and 22, which is 21.
- Values less than 21 will be classified as "No".
- Values greater than or equal to 21 will be classified as "Yes."