Quantitative Reasoning II

Prediction and Algorithms

Quantitative Concepts from Spiegelhalter Ch 6.

Algorithms, Analytics, and Prediction

Algorithms

What is an algorithm?

Definition: An algorithm is a mechanistic formula or step-by-step procedure designed to solve a specific problem or make a decision based on data.

Key Characteristics:

- Automates decision-making with minimal human intervention.
- Can be designed for practical tasks (e.g., predicting outcomes, categorizing data).

Importance: Algorithms simplify complex decisions and allow scalability across applications.

Spiegelhalter Question:

According to the reading, how do algorithms differ from scientific models?

Algorithms differ from scientific models in their purpose and approach.

Algorithms are designed to solve practical problems by using past data to produce answers with minimal human intervention. They focus on producing accurate outputs for specific tasks, such as classification or prediction, and are judged by their performance on these tasks.

In contrast, scientific models aim to understand underlying processes and explain how the world works. They prioritize gaining insights into causal relationships and are used to interpret variability and error in data.

Algorithm Types: Prediction vs Classification

How do prediction and classification differ in terms of output and purpose?

Definition of Prediction:

- Predicts future outcomes or values based on current and past data.
- Example: Forecasting stock prices or weather.

Definition of Classification:

- Assigns data into predefined categories or classes.
- Example: Determining if an email is spam or not.

Key Differences:

- Prediction focuses on estimating unknown future values, while classification involves sorting into existing categories.

Spiegelhalter Question:

In the context of predicting Titanic survival, explain how the task would differ if approached as a prediction problem versus a classification problem?

What data would you use to make predictions and classifications?

Prediction problem: the task would involve estimating the survival probability (e.g., a passenger has a 70% chance of surviving) based on features like age, gender, and ticket class (Figure 6.2). This provides a continuous output (a probability).

Classification problem: the task would involve categorizing each passenger as either a survivor or non-survivor based on these features. This results in a discrete output (e.g., survivor or non-survivor). Prediction focuses on estimating probabilities, while classification assigns data to predefined categories.

Critical Thinking Question

How could the social media app TikTok be using predictive and classification algorithms?

Machine Learning (ML) Algorithms: Supervised vs Unsupervised Learning

What is the main difference between supervised and unsupervised learning?

Supervised Learning:

- Uses labeled data where inputs are paired with correct outputs.
- The algorithm learns to predict the output from the input data features.

Unsupervised Learning

- Uses unlabeled data to identify patterns or structures.
- The algorithm explores the input data features to find clusters or associations.

Which type of ML algorithm is appropriate?

Supervised or Unsupervised

In a sprawling metropolitan hospital, doctors begin to notice an unusual pattern among patients coming in with flu-like symptoms. Many complain of high fever, a persistent dry cough, and a strange rash appearing on their upper bodies. While these symptoms overlap with common viral infections, a few doctors sense something is off. The illness appears more severe in some patients, progressing rapidly to respiratory distress without clear risk factors.

As the number of cases grows, reports emerge from other hospitals across the region. Some patients report attending the same public events or traveling through the same airport, but the connections are unclear. Public health authorities issue alerts, urging caution but stopping short of confirming an outbreak.

The hospital's research team is tasked with investigating the phenomenon. They collect an extensive dataset, including:

- **Symptoms**: A mix of flu-like signs, rashes, and respiratory complications.
- **Lab Results**: Abnormal inflammatory markers, unusual white blood cell counts, and other anomalies.
- **Patient Backgrounds**: Demographics, travel history, and potential exposure to environmental factors.
- **Temporal and Geographic Data**: Time of symptom onset and locations visited by patients.

What type of algorithm should be used investigate whether there is a novel viral outbreak?

What sort of patterns would you design the algorithm to look for?

Which type of ML algorithm is appropriate?

Supervised or Unsupervised

Possible patterns:

1. Identifying groups of patients with similar symptoms, lab results, or exposure histories.

2. Detecting geographic clusters of cases, potentially linked to common sources.

3. Uncovering trends in how the illness progresses in different individuals.

What sort of patterns would you design the algorithm to look for?

Which type of ML algorithm is this?

Supervised or Unsupervised

A tech company has been receiving complaints from users about an increase in unwanted and suspicious emails. These emails clutter inboxes with irrelevant advertisements, phishing links, and other unwanted content. To address this issue, the company's engineering team decides to develop a **spam filter** to automatically detect and block spam emails.

The team collects a large dataset of emails. Each email in the dataset is either marked as "spam" or "not spam" based on user feedback. Using this dataset, they aim to build a machine learning algorithm that can classify incoming emails as either spam or not spam, helping users keep their inboxes clean and secure.

What features or attributes of the emails might the machine learning algorithm analyze to make accurate predictions?

Which type of ML algorithm is this?

Supervised or Unsupervised

Text Content:

- Frequency of certain keywords (e.g., "free," "win," "click here").
- Presence of phrases commonly found in spam emails.

Sender Information:

- Domain name or email address of the sender.
- Known blacklisted or suspicious email domains.

Email Structure:

- Use of excessive capitalization or exclamation marks.
- Unusual formatting (e.g., hidden text, large images).

Attachments and Links:

- Presence of suspicious file types or links.
- Links directing to known malicious websites.

Metadata:

- Timestamp of when the email was sent.
- IP address of the sender.

What features or attributes of the emails might the machine learning algorithm analyze to make accurate predictions?

Training vs Test Datasets

What is the difference between training and test datasets?

Training Datasets

- Used to teach the machine learning algorithm.
- The algorithm creates a model that is used to make future classifications or predictions based on patterns or relationships.
- Ex: Spam filter: Emails labeled as “spam” or “not spam”

Test Datasets

- Used to evaluate the performance of a trained model.
- Helps determine the accuracy and generalization of the model.
- Ex: Spam filter: Contains new, unseen emails that the model has not encountered.

How to use training and test datasets

Example: Spam Filter

1. Collect Email Dataset
2. Preprocess and Label Data
3. Split Data (Training & Test)
4. Train the Model (Using Training Data)
5. Test the Model (Using Test Data)
6. Evaluate Performance
7. Deploy and Monitor

What went wrong during model training?

Scenario: Predicting Employee Turnover

A large corporation is facing high employee turnover rates and decides to build a machine learning model to predict which employees are likely to leave the company. The company gathers a dataset with detailed employee information, including job roles, salaries, performance reviews, working hours, and other metrics.

The team trains a machine learning model to identify patterns that could predict whether an employee will resign within the next year. The model performs exceptionally well during testing, achieving **99% accuracy** on the training data. Confident in its success, the company deploys the model and begins using its predictions to implement retention strategies.

However, within months of deployment, the company notices several issues:

1. The model consistently flags employees as likely to leave, even though they stay with the company.
2. Many employees who do end up leaving are not flagged by the model at all.
3. The company's retention strategies fail to reduce turnover rates.

Upon further investigation, the data science team discovers that:

- The model has memorized specific details about the training data rather than learning generalizable patterns.
- For example, it heavily relies on unique, irrelevant details like exact combinations of performance scores and salaries that only occurred in the training set.
- As a result, it performs poorly on new, unseen employee data.

Overfitting!

Definition:

- Overfitting happens when a model captures noise or randomness in the training data instead of the underlying signal.
- It occurs when the model is too complex and adapts excessively to the training data.

Key Signs of Overfitting:

- High accuracy on training data but poor performance on test data.
- Complex models with excessive parameters or rules.

Communicating Model Performance

A statistician has trained a model to predict whether a passenger of the Titanic will survive based on the data in Fig 6.2 in Spiegelhalter. The model has **82% accuracy** and and **0.78 sensitivity** in classification.

What is accuracy and sensitivity?

Accuracy: This is the percentage of total predictions (both survivors and non-survivors) that the model correctly classified. For example, if 82% accuracy is reported, it means 82% of the passengers' survival statuses (survived or not) were predicted correctly by the algorithm.

Sensitivity: This measures the proportion of actual survivors that the model correctly identified as survivors. A sensitivity of 0.78 (or 78%) indicates that the model correctly predicted 78% of the passengers who survived the Titanic disaster as survivors. It focuses on reducing false negatives (survivors incorrectly classified as non-survivors).

Quantifying Model Performance

Error Matrix: Used to quantify true positives, false positives, true negatives, and false negatives for sensitivity and specificity calculations.

Predicted Values	Actual Values	
	True Positives	False Positives
	False Negatives	True Negatives

Sensitivity: Proportion of true positives correctly identified by a model.

$$\text{Sensitivity} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

Specificity: Proportion of true negatives correctly identified by a model.

$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)} + \text{False Positives (FP)}}$$

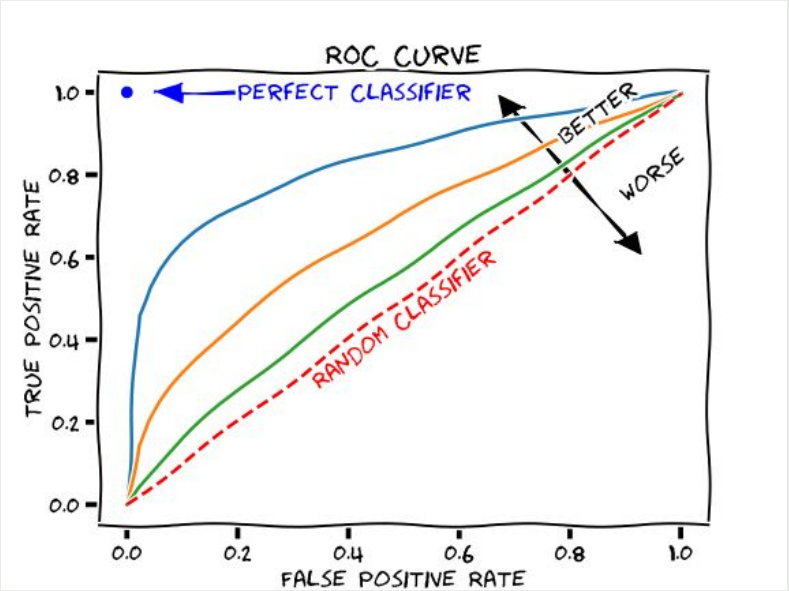
What is an ROC curve?

Receiver Operating Characteristic (ROC): originally developed during WWII to analyze radar signals.

Applied to algorithms that give a probability rather than a simple classification.

Help you identify the algorithm with the highest true positive and lowest false negative rate and the optimal amount of model training.

sensitivity



1-specificity