Module 26 – High-Level Summary and Conclusion

DSC 40A, Summer 2023

Agenda

- High-level summary of the course.
- Conclusion.

What was this course about?

Supervised Learning

The "learning from data" recipe to make predictions:

1. Choose a **prediction rule**. We've seen a few:
   - Constant: $H(x) = h$.
   - Simple linear: $H(x) = w_0 + w_1 x$.
   - Multiple linear: $H(x) = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \ldots + w_d x^{(d)}$.
2. Choose a **loss function**. others: huber loss → combination of MAE and MSE
   - Absolute loss: $L(h, y) = |y - h|$. Gamma loss → skewed target variable
   - Squared loss: $L(h, y) = (y - h)^2$. Poisson loss → skewed counts
   - 0-1 loss, UCSD loss, etc. never use these → tweedie aka compound gamma poisson
3. Minimize **empirical risk** to find optimal parameters.
   - Closed-form solutions. usually not in practice
   - Gradient descent. most common
4. Feature Engineering

huge area for tabular in particular

Unsupervised Learning

- We discussed **k-Means Clustering**, an **unsupervised machine learning** method.
  - Supervised learning: there is a "right answer" that we are trying to predict.
  - Unsupervised learning: there is no right answer, instead we're trying to find patterns in the structure of the data.

- Segmentation customers, etc
- clusters can be used as features
- label creation

Probability fundamentals

- If all outcomes in the **sample space** S are equally likely, then $P(A) = \frac{|A|}{|S|}$.
- $\bar{A}$ is the **complement** of event A. $P(\bar{A}) = 1 - P(A)$.
- Two events A, B are **mutually exclusive** if they share no outcomes, i.e. they don't overlap. In this case, the probability that A happens or B happens is $P(A \cup B) = P(A) + P(B)$.
- More generally, for any two events, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- The probability that events A and B both happen is $P(A \cap B) = P(A)P(B|A)$.
  - $P(B|A)$ is the probability that B happens given that you know A happened.
  - Through re-arranging, we see that $P(B|A) = \frac{P(A \cap B)}{P(A)}$.

Combinatorics

- A **sequence** is obtained by selecting k elements from a group of n possible elements with replacement, such that order matters.
  - Number of sequences: n^k.

- A **permutation** is obtained by selecting k elements from a group of n possible elements without replacement, such that order matters.
  - Number of permutations: $P(n, k) = \frac{n!}{(n-k)!}$.

- A **combination** is obtained by selecting k elements from a group of n possible elements without replacement, such that order does not matter.
  - Number of combinations: $\binom{n}{k} = \frac{n!}{(n-k)!k!}$.

The law of total probability and Bayes' theorem

- A set of events $E_1, E_2, \ldots, E_k$ is a **partition** of S if each outcome in S is in exactly one E_i.
- The **law of total probability** states that if A is an event and $E_1, E_2, \ldots, E_k$ is a partition of S, then

$$\begin{aligned} P(A) &= P(E_1) \cdot P(A|E_1) + P(E_2) \cdot P(A|E_2) + \ldots + P(E_k) \cdot P(A|E_k) \\ &= \sum_{i=1}^{k} P(E_i) \cdot P(A|E_i) \end{aligned}$$

- **Bayes' theorem** states that

$$P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)}$$

- We often re-write the denominator $P(A)$ in Bayes' theorem using the law of total probability.

Independence and conditional independence

- Two events A and B are **independent** when knowledge of one event does not change the probability of the other event.
  - Equivalent conditions: $P(B|A) = P(B)$, $P(A|B) = P(A)$, $P(A \cap B) = P(A) \cdot P(B)$.
- Two events A and B are **conditionally independent** if they are independent given knowledge of a third event, C.
  - Condition: $P((A \cap B)|C) = P(A|C) \cdot P(B|C)$.

Naive Bayes

- In classification, our goal is to predict a discrete category, called a **class**, given some features.
- The **Naive Bayes** classifier works by estimating the numerator of $P(\text{class}|\text{features})$ for all possible classes.
- It uses Bayes' theorem:

$$P(\text{class}|\text{features}) = \frac{P(\text{class}) \cdot P(\text{features}|\text{class})}{P(\text{features})}$$

- It also uses a "naive" simplifying assumption, that **features are conditionally independent given a class**:

$$P(\text{features}|\text{class}) = P(\text{feature}_1|\text{class}) \cdot P(\text{feature}_2|\text{class}) \cdot \ldots$$

+ smoothing

Classification Evaluation Metrics

- **Accuracy:** Ratio of correctly predicted observations to the total observations.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

- **Precision:** Ratio of correctly predicted positive observations to the total predicted positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- **Recall:** Ratio of correctly predicted positive observations to the all observations in actual class.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- **F1 Score:** Harmonic mean of Precision and Recall.

Classification Loss Functions (Mentioned During Discussion)

- **Log Loss:** Logarithm of the likelihood of the true label given the predicted probabilities.

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

- Also known as cross-entropy loss.

- Many other options for different situations.

  - focal loss

Conclusion

Learning objectives

At the start of the quarter, we told you that by the end of DSC 40A, you'll...

- understand the basic principles underlying almost every machine learning and data science method.
- be able to tackle problems such as:
  - How do we know if an avocado is going to be ripe before we eat it?
  - How do we teach a computer to read handwritten text?
  - How do we predict a future data scientist's salary?

What's next?

In DSC 40A, we just scratched the surface of the theory behind data science. In future courses, you'll build upon your knowledge from DSC 40A, and will learn:

- More supervised learning.
  - Logistic regression, decision trees, neural networks, etc.
- More unsupervised learning.
  - Other clustering techniques, PCA, etc.
- More probability.
  - Random variables, distributions, etc.
- More connections between all of these areas.
  - For instance, you'll learn how probability is related to linear regression.
- More practical tools.

Note on grades

- Grades do not define you.
- Interview committees will be much more interested in skills and portfolio. → how well do you interview?
  – coding, ML, DS
- Graduate admission committees are more interested in research potential.
- Learning does not end at university

Thank you!

- This course would not have been possible without our TAs Fatemeh and Anna.
- It also would not have been possible without our 3 TAs Daniel, Vivian, and Yujia.