Module 11: Naïve Bayes

Learning outcomes

1. Evaluate the feasibility and limitations of using exact Bayes and naïve Bayes approaches.
2. Apply Bayes' theorem to calculate probabilities.
3. Apply the naïve Bayes algorithm to classification tasks in theory and practice.
4. Analyse Laplace smoothing in naïve Bayes classification with sparse data.
5. Use binning methods to convert numerical features into categorical data.

Bayes' theorem

- Bayes' theorem lets you update a probability estimate based on new information.
- This forms the foundation for both exact Bayes and naïve Bayes classifiers.
- The formula is: $P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$

  where:
  - $\mathrm{P(A \mid B)}$ is the posterior probability
  - $\mathrm{P(B \mid A)}$ is the likelihood
  - $\mathrm{P(A)}$ is the prior probability
  - $\mathrm{P(B)}$ is the marginal likelihood

Exact Bayes classifier

- A conceptual classifier that looks for exact matches in the training data
- Intuitive but impractical when inputs are high-dimensional
- Prediction formula: $P(C_k|X) = \frac{P(X|C_k) \times P(C_k)}{P(X)}$

  where:
  - $P(X|C_k)$ is the frequency of X in class C_k
  - $P(C_k)$ is the prior probability of class C_k
  - $P(X)$ is the overall frequency of X in the data set

Naïve Bayes classifier

- A practical version that assumes class-conditional independence between features
- The naïve assumption:
  $P(F_1, F_2, \ldots, F_n|C) = P(F_1|C) \times P(F_2|C) \ldots \times P(F_n|C)$
- The naïve Bayes formula:
  $P(C|F_1, F_2, \ldots, F_n) = \frac{P(F_1|C) \times P(F_2|C) \ldots \times P(F_n|C) \times P(C)}{P(F_1, F_2, \ldots, F_n)}$

  where:
  - $P(C|F)$ is the posterior probability of class C given features F
  - $P(C)$ is the prior probability of class C
  - $P(F|C)$ is the likelihood of features F given the class C

  - $P(F)$ is the total probability of features
- The classification rule:
  $Predictedclass = argmax_C P(C|\mathrm{F}_1, \mathrm{F}_2, \ldots, \mathrm{F}_\mathrm{n})$

- Laplacian smoothing:

$$P(F_i \ = \ x|C) \ = \frac{N_{x|C} \ + \ k}{N_c \ + \ k \times N}$$

  where:
  - $N_{x|C}$ is the count of feature value x in class C
  - N_C is the total count of all feature values in class C
  - N is the number of possible values the feature can take
  - Parameter k is the smoothing parameter

Class-conditional independence:

- Assumes features are independent within each class
- Reduces probability estimates from exponential to linear
- Enables effective learning from limited data
- Must be applied after Bayes' theorem to keep probabilities valid

The Laplace estimator:

- Adds one to all counts to avoid zero probabilities
- Ensures unseen features don't zero out class predictions
- Helps the model generalise better to new data
- Prevents overfitting and keeps probability estimates valid

Converting numeric features to categorical

- **Manual binning:** use domain knowledge or standard conventions to define bins.
- **Equal-width binning:** divide the full numeric range into equal-sized intervals.
- **Equal-frequency (quantile) binning:** divide the data into bins with approximately equal numbers of observations.