Machine Learning (WiSe 2025/2026)

Author: Suvansh Shukla
Matriculation No. 256245

Assignment 3 Task 3.2

(A) These are the various formulae used in Naive Bayes Classifier:

$P(C_k|X) = \frac{P(X|C_k).P(C_k)}{p(X)}$

In classification we don't need the exact probability, we only need the most likely class so we want the max:

$\hat{C} = argmax P(C_k|X)$

Then substitute Bayes' theorem,

$\hat{C} = argmax \frac{P(X)P(X|C_k)}{P(C_k)}$

Now since $P(X)$ is constant across all classes and doesn't depend on C_k we can drop it

$\hat{C} = argmax P(X|C_k)P(C_k)$

This is the **Maximum A Posteriori (MAP) decision rule**.
The class we predict is the one with the highest *posterior probability* given the observed features.

(B) The naive assumption made here is that all features are independent and do not affect each other.
This is important because it works surprisingly well and can help us estimate ratios of probabilities, even though they may not be independent.
This is useful in contexts where features are very weakly correlated, like words in an email.
Though in reality this assumption is rarely true as most the occurrences of things are linked.
Thus, this means that Naive Bayes performs poorly when features are clearly highly correlated, e.g. pixels in an image.

(C) The probabilities in a Naive Bayes are calculated by taking the number of occurrences of an instance and dividing it by the

total number of instances.
Then it takes into account how probable the occurrence of one event is given that another event has already occurred.
Mathematically, it is represented like so:

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

Where,
P(A|B) is the posterior probability
P(B|A) is the Likelihood
P(A) is the prior probability
P(B) is the marginal probability of the evidence

(D) When a feature never appears in a class (zero frequency problem), we can use Laplace smoothing.
This is done by pretending we've seen each possible feature value at least once or (α-times).
Mathematically represented like this:

$$P(x_i|C_k) = \frac{count(x_i,C_k)+\alpha}{\sum_x(count((x',C_k)+\alpha)}$$

Missing values we must be handled in different ways.

- Imputation: replace missing values with estimated values
- Add "Missing" as its own valid category
- summing or integrating over unknown variables (marginalization), this is sort of like taking all combinations of the missing value into account