Tutorial: Naive Bayes

Cheat Sheet and Practice Problems

ES335 - Machine Learning
IIT Gandhinagar

July 23, 2025

1 Summary from Slides

1.1 Bayes' Theorem Foundation

Bayes' Theorem:

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

For Machine Learning Classification:

$$P(y|x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n|y)P(y)}{P(x_1, x_2, \ldots, x_n)}$$

Where:

- $P(y|x_1, x_2, \ldots, x_n)$: Posterior probability (what we want to predict)
- $P(x_1, x_2, \ldots, x_n|y)$: Likelihood of features given class
- $P(y)$: Prior probability of class
- $P(x_1, x_2, \ldots, x_n)$: Evidence (normalizing constant)

1.2 The Naive Assumption

Why "Naive"? Assumes features are conditionally independent given the class:

$$P(x_1, x_2, \ldots, x_n|y) = P(x_1|y) \cdot P(x_2|y) \cdots P(x_n|y) = \prod_{i=1}^{n} P(x_i|y)$$

This simplifies the model dramatically but is often violated in practice.

1.3 Naive Bayes Classification Rule

Prediction: Choose class with highest posterior probability

$$\hat{y} = \arg\max_y P(y|x_1, x_2, \ldots, x_n) = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

Since denominator $P(x_1, x_2, \ldots, x_n)$ is constant across classes, we can ignore it.

1.4 Types of Naive Bayes

1. Categorical/Multinomial Naive Bayes (for discrete features):

$$P(x_i = k|y = c) = \frac{\text{Count}(x_i = k \text{ and } y = c)}{\text{Count}(y = c)}$$

2. Gaussian Naive Bayes (for continuous features):

$$P(x_i = v|y = c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(v - \mu_c)^2}{2\sigma_c^2}\right)$$

Where μ_c and σ_c^2 are the mean and variance of feature x_i for class c.

1.5 Parameter Estimation

Prior Probabilities:

$$P(y = c) = \frac{\text{Count}(y = c)}{N}$$

Likelihood for Categorical Features:

$$P(x_i = k|y = c) = \frac{\text{Count}(x_i = k, y = c)}{\text{Count}(y = c)}$$

Likelihood for Gaussian Features:

$$\mu_c = \frac{1}{N_c} \sum_{i:y_i=c} x_i, \quad \sigma_c^2 = \frac{1}{N_c} \sum_{i:y_i=c} (x_i - \mu_c)^2$$

1.6 Laplace Smoothing

Problem: Zero probabilities when feature values not seen in training data.
Solution: Add pseudocounts ($\alpha = 1$ for Laplace smoothing):

$$P(x_i = k|y = c) = \frac{\text{Count}(x_i = k, y = c) + \alpha}{\text{Count}(y = c) + \alpha \cdot |\text{vocabulary}|}$$

2 Practice Problems

Problem : Basic Bayes' Theorem

A medical test for a rare disease has the following characteristics:

- $P(\text{Test} = +|\text{Disease} = \text{True}) = 0.99$
- $P(\text{Test} = -|\text{Disease} = \text{False}) = 0.99$
- $P(\text{Disease} = \text{True}) = 0.0001$

If someone tests positive, what is the probability they actually have the disease?

Problem : Spam Email Classification - Setup

Given the following email dataset with binary features (word present = 1, absent = 0):

Email	"free"	"money"	Class
1	1	0	Spam
2	1	1	Spam
3	0	1	Spam
4	0	0	Ham
5	1	0	Ham
6	0	0	Ham

Calculate all prior probabilities $P(\text{Spam})$ and $P(\text{Ham})$.

Problem : Likelihood Calculation

Using the spam email dataset from Problem 2, calculate:

- $P(\text{"free"} = 1|\text{Spam})$
- $P(\text{"free"} = 0|\text{Spam})$
- $P(\text{"money"} = 1|\text{Ham})$
- $P(\text{"money"} = 0|\text{Ham})$

Problem : Classification Decision

Using the spam email dataset, classify a new email with features ["free" = 1, "money" = 0]. Show all calculations and determine whether it's Spam or Ham.

Problem : Gaussian Naive Bayes Setup

For the height/weight/gender classification problem:

Height	Weight	Foot Size	Gender
6.0	180	12	M
5.92	190	11	M
5.58	170	12	M
5.92	165	10	M
5.0	100	6	F
5.5	150	8	F
5.42	130	7	F
5.75	150	9	F

Calculate the mean and variance for height for both Male and Female classes.

Problem : Gaussian Probability Calculation

Using the Gaussian parameters from Problem 5, calculate $P(\text{Height} = 6.0|\text{Male})$ using the Gaussian probability density function.

Problem : Complete Gaussian Classification

Classify a person with Height = 6.0, Weight = 130, Foot Size = 8 using Gaussian Naive Bayes. Calculate the posterior probabilities for both Male and Female classes.

Problem : Laplace Smoothing

Consider a text classification problem with vocabulary = ["good", "bad", "movie"]:
Training data:

- Positive: ["good movie"], ["good"]
- Negative: ["bad movie"], ["bad"]

Calculate $P(\text{"excellent"}|\text{Positive})$ with and without Laplace smoothing ($\alpha = 1$).

Problem : Independence Assumption Analysis

Consider features: "Age" and "Income" for credit approval.

- Real correlation between Age and Income: 0.7
- Naive Bayes assumes: correlation = 0

Explain how this independence assumption might affect: a) Model accuracy b) Feature importance interpretation c) When the model might still work well despite violated assumptions

Problem : Multinomial Naive Bayes

For document classification, you have word counts:
Document 1 (Sports): "game": 3, "team": 2, "player": 1 Document 2 (Politics): "government": 2, "policy": 3, "vote": 1
Calculate the probability of the word "team" given the Sports class using multinomial distribution parameters.

Problem : Log-Space Computation

When multiplying many small probabilities, we use log-space to avoid numerical underflow.
Transform this calculation to log-space:

$$P(\text{Class}|\text{features}) = P(\text{Class}) \times P(f_1|\text{Class}) \times P(f_2|\text{Class}) \times P(f_3|\text{Class})$$

Given: $P(\text{Class}) = 0.3$, $P(f_1|\text{Class}) = 0.1$, $P(f_2|\text{Class}) = 0.05$, $P(f_3|\text{Class}) = 0.02$

Problem : Handling Missing Features

In the spam classification dataset, suppose a new email has:

- "free": present
- "money": missing (unknown)
- "urgent": not in original vocabulary

Describe three strategies for handling missing and unknown features in Naive Bayes classification.

Problem : Model Evaluation

You trained a Naive Bayes classifier with the following confusion matrix:

	Predicted: +	Predicted: -
Actual: +	80	20
Actual: -	10	90

Calculate: a) Accuracy b) Precision for positive class c) Recall for positive class d) F1-score

Problem : Feature Selection Impact

You have 1000 features for text classification. When you:

- Use all 1000 features: Accuracy = 85%
- Use top 100 features (by mutual information): Accuracy = 88%
- Use top 50 features: Accuracy = 84%

Explain why reducing features can improve Naive Bayes performance. What does this suggest about the independence assumption?

Problem : Real-world Application Design

Design a Naive Bayes system for movie review sentiment analysis:
Dataset: 10,000 movie reviews (positive/negative) **Features**: Word frequencies, review length, rating
a) Choose appropriate Naive Bayes variant for each feature type b) Describe preprocessing steps c) How would you handle the class imbalance (70% positive, 30% negative)? d) Design evaluation methodology e) Identify potential limitations and when Naive Bayes might fail