Intermediate Data Science

Naive Bayes Classifiers

Joanna Bieri DATA201

Mini Review - Naive Bayes Classifiers

REVIEW

- Naive Bayes classifiers are fairly simple and have strong assumptions but they can perform surprisingly well for high dimensional data.
- They work well with text data: spam detection, sentiment analysis, document classification.
- They are fast, need little data, and make strong baselines.

Mini Review - How It Works

REVIEW

- **Start with prior knowledge.** *Example:* Maybe 80% of emails are "not spam" and 20% are "spam."
- **Look at the evidence.** The model checks which features appear (like certain words in an email).
- **Combine the clues.** It assumes that all clues are independent - that's why it's called **Naive** Bayes.
- **Pick the most likely class.**

Mini Review - Bayes' Theorem

REVIEW

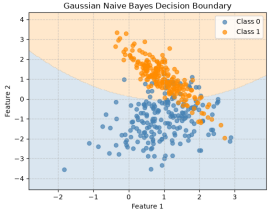
$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}$$

- $P(C_k|x)$: posterior probability of class C_k given input x
- $P(x|C_k)$: likelihood of the data given the class, $P(C_k)$: prior probability of the class

The "naive" assumption: $P(x|C_k) = \prod_{i=1}^{n} P(x_i|C_k)$

Mini Review - Gaussian Naive Bayes

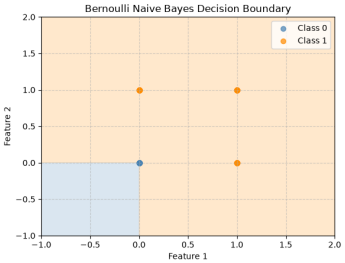
REVIEW

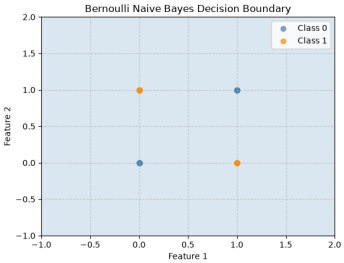



```
gnb = GaussianNB()
gnb.fit(X_train, y_train)
y_pred = gnb.predict(X_test)
```

For continuous features that are roughly normally distributed.

Mini Review - Bernoulli Naive Bayes: OR and XOR

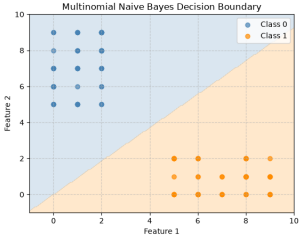




For binary (0/1) features. OR: accuracy 1.0. XOR: accuracy 0.52, the class depends on the interaction between features.

Mini Review - Multinomial Naive Bayes

REVIEW



Best used for data where the features represent **discrete counts or frequencies**, such as how many times a word appears in a document.

Mini Review - Application: Spam Collection

REVIEW

- **Clean and normalize text**: lower case, remove punctuation and stopwords, stemming.
- **Convert the text into vectors**: `TfidfVectorizer` down-weights words that appear in many documents.
- **Train and test our classifier**: `MultinomialNB`, 42 wrong out of 1115 test messages.

Naive Bayes

ASK

In your own words:
why is Naive Bayes
called "naive"?

Naive Bayes

ANSWER

In your own words: why is Naive Bayes called "naive"?

- It assumes every feature is independent of the others, once you know the class.
- So it can just multiply the probability of each clue.
- That is almost never true in real data, and it often works anyway.

Naive Bayes

ASK

In a spam filter,

what is the prior?

What is the evidence?

Naive Bayes

ANSWER

In a spam filter, what is the prior? What is the evidence?

- The prior: how common spam is before you read the email, eg. 20% of emails are spam.
- The evidence: the words in this email, and how often each word shows up in spam and in normal email.
- Bayes' theorem turns them into the chance THIS email is spam.

Which Naive Bayes?

ASK

Which Naive Bayes would you use for:

(1) heights and weights,

(2) did the customer click: yes or no,

(3) how many times each word appears?

Which Naive Bayes?

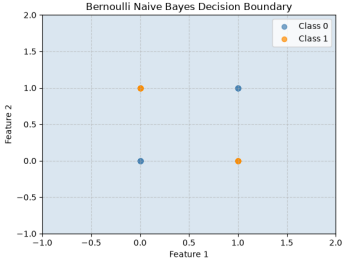
ANSWER

Which Naive Bayes: heights and weights, clicked yes or no, word counts?

- Heights and weights are continuous: **Gaussian**.
- Yes or no (0 or 1): **Bernoulli**.
- Counts: **Multinomial**.

Which Naive Bayes?

ASK



This is the XOR data. Why can Naive Bayes not get it right?

Which Naive Bayes?

ANSWER

This is the XOR data. Why can Naive Bayes not get it right?

- In XOR the class depends on the two features together.
- Each feature by itself says nothing about the class, and Naive Bayes only looks at one feature at a time.
- It got 0.52, about a coin flip.

Spam Detection

ASK

Machine learning models need numbers.

What does TF-IDF do to a word

like "the" that is in every message?

Spam Detection

ANSWER

What does TF-IDF do to a word like "the" that is in every message?

- It turns each message into a vector of numbers, one per word.
- A word that appears in many messages gets down-weighted, so "the" counts for almost nothing.
- A word that is common in a few messages, like "free" or "claim", counts a lot.

Spam Detection

ASK

`spam` row of the report

	precision	recall
spam	1.00	0.72

Which mistake is the model making? Is that the mistake you want a spam filter to make?

Spam Detection

ANSWER

Spam precision 1.00, recall 0.72. Which mistake is the model making? Is it the one you want?

- Precision 1.00: every message it called spam really was spam.
- Recall 0.72: some spam slipped through to the inbox.
- That is the better mistake. A real message from a friend in the spam folder costs more.

INTERVIEW

What is the difference between the frequentist and the Bayesian approaches to probability?

INTERVIEW

Define precision, recall and F1.

What is the trade-off between them?

Questions?