A crash course in probability and Naive Bayes classification

Chapter 9

Probability theory

Random variable: a variable whose possible values are numerical outcomes of a random phenomenon.

Examples: A person’s height, the outcome of a coin toss

Distinguish between discrete and continuous variables.

The distribution of a discrete random variable:
The probabilities of each value it can take.
Notation: \( P(X = x_i) \).
These numbers satisfy:

\[
\sum_i P(X = x_i) = 1
\]

Probability theory

Marginal Probability

\[
p(X = x_i) = \frac{c_i}{N},
\]

Joint Probability

\[
p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}
\]

Conditional Probability

\[
p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}
\]

A joint probability distribution for two variables is a table.

If the two variables are binary, how many parameters does it have?

What about joint probability of \( d \) variables \( P(X_1, ..., X_d) \)?

How many parameters does it have if each variable is binary?
Probability theory

Marginalization:
\[ p(X = x_i) = \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^{L} n_{ij} \]
\[ p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \]
\[ p(Y = y_j | X = x_i) p(X = x_i) \]

Product Rule
\[ p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \]
\[ p(Y = y_j | X = x_i) p(X = x_i) \]

Using probability in learning

Interested in: \( P(Y | X) \)

\( P(Y = \ast | X) \)

\( P(Y = \ast | X) \)

For example, when classifying spam, we could estimate \( P(Y | X) \) Viagura, lottery)

We would then classify an example if \( P(Y | X) > 0.5 \).

However, it’s usually easier to model \( P(X | Y) \)

The Rules of Probability

Marginalization
\[ p(X) = \sum_{Y} p(X,Y) \]

Product Rule
\[ p(X,Y) = p(Y|X)p(X) \]

Independence: \( X \) and \( Y \) are independent if \( P(Y|X) = P(Y) \)

This implies \( P(X,Y) = P(X)P(Y) \)

Maximum likelihood

Fit a probabilistic model \( P(x | \Theta) \) to data

- Estimate \( \Theta \)

Given independent identically distributed (i.i.d.) data \( X = (x_1, x_2, ..., x_n) \)

- Likelihood
\[ P(X | \Theta) = P(x_1 | \Theta)P(x_2 | \Theta), ..., P(x_n | \Theta) \]

- Log likelihood
\[ \ln P(X | \Theta) = \sum_{i=1}^{n} \ln P(x_i | \Theta) \]

Maximum likelihood solution: parameters \( \Theta \) that maximize \( \ln P(X | \Theta) \)
Example

Example: coin toss
Estimate the probability \( p \) that a coin lands "Heads" using the result of \( n \) coin tosses, \( h \) of which resulted in heads.

The likelihood of the data:

\[
P(X|\theta) = p^h (1-p)^{n-h}
\]

Log likelihood:

\[
\ln P(X|\theta) = h \ln p + (n-h) \ln(1-p)
\]

Taking a derivative and setting to 0:

\[
\frac{\partial \ln P(X|\theta)}{\partial p} = \frac{h}{p} - \frac{(n-h)}{(1-p)} = 0
\]

\[
\Rightarrow p = \frac{h}{n}
\]

Bayes’ rule

From the product rule:

\[
P(Y, X) = P(Y|X) P(X)
\]

and:

\[
P(Y, X) = P(X|Y) P(Y)
\]

Therefore:

\[
P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}
\]

This is known as Bayes’ rule

Maximum a-posteriori and maximum likelihood

The maximum a posteriori (MAP) rule:

\[
y_{MAP} = \arg \max_y P(Y|X) = \arg \max_y \frac{P(X|Y)P(Y)}{P(X)} = \arg \max_y P(X|Y)P(Y)
\]

If we ignore the prior distribution or assume it is uniform we obtain the maximum likelihood rule:

\[
y_{ML} = \arg \max_y P(X|Y)
\]

A classifier that has access to \( P(Y|X) \) is a Bayes optimal classifier.
Naive Bayes classifier

We would like to model \( P(X \mid Y) \), where \( X \) is a feature vector, and \( Y \) is its associated label.

Task: Predict whether or not a picnic spot is enjoyable.

Training Data: \( X = (X_1, X_2, X_3, \ldots, X_d) \) \( Y \)

<table>
<thead>
<tr>
<th>Sky</th>
<th>Temp</th>
<th>Humid</th>
<th>Wind</th>
<th>Water</th>
<th>Forecast</th>
<th>EnjoySpt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Cold</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Change</td>
<td>No</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Cool</td>
<td>Change</td>
<td>Yes</td>
</tr>
</tbody>
</table>

How many parameters?
Prior: \( P(Y) \) \( k-1 \) if \( k \) classes
Likelihood: \( P(X \mid Y) \) \( 2^d - 1 \)k for binary features

Naive Bayes classifier

We would like to model \( P(X \mid Y) \), where \( X \) is a feature vector, and \( Y \) is its associated label.

Simplifying assumption: conditional independence: given the class label the features are independent, i.e.

\[
P(X \mid Y) = P(x_1 \mid Y)P(x_2 \mid Y)\ldots P(x_d \mid Y)
\]

How many parameters now? \( dk + k - 1 \)

Naive Bayes classifier

Naive Bayes decision rule:

\[
y_{NB} = \arg \max_Y P(X \mid Y)P(Y) = \arg \max_Y \prod_{i=1}^{d} P(x_i \mid Y)P(Y)
\]

If conditional independence holds, NB is an optimal classifier!
Training a Naïve Bayes classifier

Training data: Feature matrix $X$ (n x d) and labels $y_1$–$y_n$

Maximum likelihood estimates:

Class prior: $\hat{P}(y) = \frac{\{i: y_i = y\}}{n}$

Likelihood: $\hat{P}(x_i | y) = \frac{\hat{P}(x_i, y)}{P(y)} = \frac{\{i: X_{ij} = x_i, y_i = y\}/n}{\{i: y_i = y\}/n}$

Example

Email classification: training data

<table>
<thead>
<tr>
<th>E-mail</th>
<th>$a$?</th>
<th>$b$?</th>
<th>$c$?</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_1$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>$e_2$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>$e_3$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>$e_4$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>$e_5$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>−</td>
</tr>
<tr>
<td>$e_6$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>−</td>
</tr>
<tr>
<td>$e_7$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>−</td>
</tr>
<tr>
<td>$e_9$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>−</td>
</tr>
</tbody>
</table>

What are the parameters of the model?

\[
\hat{P}(y) = \frac{\{i: y_i = y\}}{n}
\]

\[
\hat{P}(x_i | y) = \frac{\hat{P}(x_i, y)}{P(y)} = \frac{\{i: X_{ij} = x_i, y_i = y\}/n}{\{i: y_i = y\}/n}
\]
Example

Email classification: training data

<table>
<thead>
<tr>
<th>Event</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>e_1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>e_2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>e_3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>e_4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>e_5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>e_6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>e_7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>e_8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>

What are the parameters of the model?

\[ P(+) = 0.5, \quad P(-) = 0.5 \]
\[ P(a|+) = 0.5, \quad P(a|-) = 0.75 \]
\[ P(b|+) = 0.75, \quad P(b|-) = 0.25 \]
\[ P(c|+) = 0.25, \quad P(c|-) = 0.25 \]

Comments on Naïve Bayes

Usually features are not conditionally independent, i.e.

\[ P(X|Y) \neq P(x_1|Y)P(x_2|Y), \ldots, P(x_d|Y) \]

And yet, one of the most widely used classifiers. Easy to train!

It often performs well even when the assumption is violated.


When there are few training examples

What if you never see a training example where \( x_1 = a \) when \( y = \text{spam} \)?

\[ P(x \mid \text{spam}) = P(a \mid \text{spam})P(b \mid \text{spam})P(c \mid \text{spam}) = 0 \]

What to do?

Add "virtual" examples for which \( x_1 = a \) when \( y = \text{spam} \).
Naïve Bayes for continuous variables

Need to talk about continuous distributions!

Continuous Probability Distributions

The probability of the random variable assuming a value within some given interval from \(x_1\) to \(x_2\) is defined to be the area under the graph of the probability density function between \(x_1\) and \(x_2\).

Expectations

Discrete variables

\[
E[f] = \sum_x p(x)f(x)
\]

Continuous variables

\[
E[f] = \int p(x)f(x) \, dx
\]

Conditional expectation (discrete)

\[
E_x[f|y] = \sum_x p(x|y)f(x)
\]

Approximate expectation (discrete and continuous)

\[
E[f] \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n)
\]

The Gaussian (normal) distribution

\[
\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ \frac{-1}{2\sigma^2} (x - \mu)^2 \right\}
\]

\[
\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \, dx = 1
\]

\[
E[x] = \int_{-\infty}^{\infty} x \mathcal{N}(x|\mu, \sigma^2) \, dx = \mu
\]

\[
\text{var}[x] = E[x^2] - E[x]^2 = \sigma^2
\]
Properties of the Gaussian distribution

Standard Normal Distribution

A random variable having a normal distribution with a mean of 0 and a standard deviation of 1 is said to have a standard normal probability distribution.

Standard Normal Probability Distribution

Converting to the Standard Normal Distribution

\[ z = \frac{x - \mu}{\sigma} \]

We can think of \( z \) as a measure of the number of standard deviations \( x \) is from \( \mu \).

Gaussian Parameter Estimation

Likelihood function

\[ p(x|\mu, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(x_i|\mu, \sigma^2) \]
Maximum (Log) Likelihood

$$\ln p(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi)$$

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \quad \sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2$$

Gaussian models

Assume we have data that belongs to three classes, and assume a likelihood that follows a Gaussian distribution

Gaussian Naïve Bayes

Likelihood function:

$$P(X_i = x | Y = y_k) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} \exp \left( -\frac{(x - \mu_{ik})^2}{2\sigma_{ik}^2} \right)$$

Need to estimate mean and variance for each feature in each class.
Summary

Naive Bayes classifier:
- What’s the assumption
- Why we make it
- How we learn it
Naive Bayes for discrete data
Gaussian naive Bayes