A softbot that finds interesting newsgroup articles

The problem of document classification is to assign a given text document into one of a set of categories.

Examples:

- Newsgroup articles on Linux.
- Web pages that pertain to dogs as pets.
- Library books that discuss hidden Markov models.
- Newspaper articles on European currency unification (ECU)

Bayes Rule

Note that since

$$P(A \mid B)P(B) = P(B \mid A)P(A)$$

we get that

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

This is the single-most important idea in all of probabilistic reasoning!

We can express it for distributions as

$$\mathbf{P}(Y \mid X) = \frac{\mathbf{P}(X \mid Y)\mathbf{P}(Y)}{\mathbf{P}(X)}$$

A Medical Diagnosis Problem

Suppose a doctor is trying to find out if a patient is suffering from some type of cancer. If the cancer is only found on average in 2 out of every 1000 people, the doctor's initial beliefs can be expressed as

$$P(cancer) = 0.002$$

$$P(\neg cancer) = 0.992$$

Conditional probabilities in Medical Diagnosis

There is a laboratory test to determine if the patient has cancer. Unfortunately, this test is not 100% accurate. The test comes back positive in 98% of cases where the patient has cancer. Also, the test comes out negative only in 97% of the cases where the patient does not have cancer.

$$P(+ \mid cancer) = 0.98 \quad P(- \mid \neg cancer) = 0.97$$

If the doctor orders a test, and it comes back positive, what is the probability that the patient indeed has cancer?

Bayes Rule

Since

$$P(A \wedge B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

we get that

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

This is the most important theorem in all of probabilistic reasoning!

Bayes Rule applied to Medical Diagnosis

We are asked to determine

$$P(cancer \mid +)$$

Using Bayes rule, we get that

$$P(cancer \mid +) = \frac{P(+ \mid cancer)P(cancer)}{P(+)}$$

We also need to find out the chance that the patient does not have cancer.

$$P(\neg cancer \mid +) = \frac{P(+ \mid \neg cancer)P(\neg cancer)}{P(+)}$$

Medical Diagnosis Problem

The data given for this problem can be put in the form of a table as shown below.

Test / Actual	**Cancer**	**¬Cancer**
+	.98	.03
-	.02	.97

Bayes Rule applied to Medical Diagnosis Problem

To determine if the patient has cancer, we have to compute

$$\begin{aligned} P(cancer \mid +) &= \frac{P(+ \mid cancer)P(cancer)}{P(+)} \\ &= \frac{0.98 \times 0.002}{P(+)} \end{aligned}$$

To determine if the patient does not has cancer, we have to compute

$$\begin{aligned} P(\neg cancer \mid +) &= \frac{P(+ \mid \neg cancer)P(\neg cancer)}{P(+)} \\ &= \frac{0.03 \times 0.992}{P(+)} \end{aligned}$$

Bayes Rule applied to Medical Diagnosis Problem

How do we estimate the denominator $P(+)$? We simply ignore it, because we are only interested in the most likely hypothesis!

So, to determine if the patient has cancer, we compute

$$P(+ \mid cancer)P(cancer) = 0.98 \times 0.002 = 0.00196$$

To determine if the patient does not has cancer, we compute

$$P(+ \mid \neg cancer)P(\neg cancer) = 0.03 \times 0.992 = 0.02976$$

Normalization of Probabilities

We can always recover the probabilities by normalizing, since we know that one of the two hypotheses must be true. That is,

$$P(cancer \mid +) \ + \ P(\neg cancer \mid +) = 1$$

So we get

$$P(cancer \mid +) = \frac{0.00196}{0.00196 + 0.02976} = 0.06$$

and similarly,

$$P(\neg cancer \mid +) = \frac{0.02976}{0.00196 + 0.02976} = 0.94$$

So, it is very unlikely that the patient has cancer!

Bayesian Learning

- Bayes rule
- MAP and ML estimators
- MDL principle
- Naive Bayes Classifer and an application to text classification

Bayes Rule

Consider a hypothesis space H and a set of training data D. We are interesed in determining the best (most probable) hypothesis $h \in H$ given the data.

Let $P(h)$ denote the *prior probability* of a hypothesis. We can set these to reflect any a priori knowledge we have regarding the likelihood of the hypotheses. We can also assume *maximum likelihood*, meaning all hypotheses are equally likely.

Bayes Rule: The posterior probability of a hypothesis h given training data (evidence) D can be computed from the a priori probability as

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

MAP and ML Hypotheses

We can use Bayes rule to determine the best hypotheses given the training data.

Maximum A Posteriori (MAP) Hypothesis is defined as follows.

$$\begin{aligned} h_{MAP} &\equiv \text{argmax}_{h \in H} P(h \mid D) \\ &= \text{argmax}_{h \in H} \frac{P(D \mid h) P(h)}{P(D)} \\ &= \text{argmax}_{h \in H} P(D \mid h) P(h) \end{aligned}$$

Maximum Likelihood (ML) Hypothesis is defined, assuming that all hypotheses are equally likely $(P(h_i) = P(h_j) \ \ \forall i, j)$.

$$\begin{aligned} h_{ML} &\equiv \text{argmax}_{h \in H} P(h \mid D) \\ &= \text{argmax}_{h \in H} \frac{P(D \mid h) P(h)}{P(D)} \\ &= \text{argmax}_{h \in H} P(D \mid h) \end{aligned}$$

Example

Consider the problem of predicting whether it will rain today (the two hypotheses are $rain$ and $\neg rain$.) From weather data in East Lansing, we might find out that

$$P(rain) = .10 \qquad\qquad P(\neg rain) = .9$$

Furthermore, we will predict rain based on whether it was cloudy yesterday. We are given that

$$\begin{aligned} P(cloudy \mid rain) = 0.9 &\qquad P(sunny \mid rain) = 0.1 \\ P(cloudy \mid \neg rain) = 0.3 &\qquad P(sunny \mid \neg rain) = 0.7 \end{aligned}$$

Suppose that yesterday was cloudy. What is the best hypotheses?

Example (contd.)

The MAP hypothesis:

$$\begin{aligned} h_{MAP} &\equiv \text{argmax}_{h\in H} P(D \mid h)P(h) \\ &= \text{argmax}_{h\in H}\,(P(c \mid rain)P(rain), P(c \mid \neg rain)P(\neg rain)) \\ &= \text{argmax}_{h\in H}\,((0.9 * 0.1), (0.3 * .9)) \\ &= \neg rain \end{aligned}$$

The ML hypothesis:

$$\begin{aligned} h_{ML} &\equiv \text{argmax}_{h\in H} P(D \mid h) \\ &= \text{argmax}_{h\in H}\,(P(cloudy \mid rain), P(cloudy \mid \neg rain)) \\ &= \text{argmax}_{h\in H}\,((0.9), (0.3)) \\ &= rain \end{aligned}$$

MAP Learning and Version Spaces

The MAP algorithm is impractical because it requires computing

$$\text{argmax}_{h\in H} P(h \mid D)$$

over the whole hypothesis space H. But it offers a useful framework for understanding concept learning programs such as version spaces.

Assuming noise-free instances, and VS bias is correct (i.e. the target concept c is correctly described by some $h \in H$), then using ML principle

$$P(h) = \frac{1}{|H|} \qquad \forall h \in H$$

Given a fixed sequence of instances $D < x_1, d_1 >, \ldots, < x_n, d_n >$. What is the MAP hypothesis h given D?

MAP Learning and Version Spaces

If we want to use Bayes rule, we first need to compute

$$\begin{aligned} P(D \mid h) &= 1 \quad \text{if} \;\; h(x_i) = d_i \;\; \forall d_i \in D \\ &= 0 \;\; \text{otherwise} \end{aligned}$$

Then, we have (assuming h is consistent with D)

$$\begin{aligned} P(h \mid D) &= \frac{P(D \mid h)P(h)}{P(D)} \\ &= \frac{1.\frac{1}{|H|}}{\frac{|VS_{H,D}|}{|H|}} \\ &= \frac{1}{|VS_{H,D}|} \end{aligned}$$

Thus, every consistent hypothesis is a MAP hypothesis. If a hypothesis is inconsistent with the data D, it is not MAP.

Minimum Description Length (MDL) Principle

Occam's razor says that a "short" hypothesis that explains the data is to be preferred over longer hypotheses. The MDL principle formalizes this heuristic, using the Bayesian framework.

$$\begin{aligned} h_{MAP} &\equiv \text{argmax}_{h \in H} P(D \mid h)P(h) \\ &= \text{argmax}_{h \in H} log_2 P(D \mid h) + log_2 P(h) \\ &= \text{argmin}_{h \in H} - log_2 P(D \mid h) - log_2 P(h) \end{aligned}$$

From basic coding theory, we know that the if a message i appears with probability p_i, the shortest code for i requires $-log_2 p_i$ bits.

Minimum Description Length (MDL) Principle

- $-log_2 P(h)$ is the description length of hypothesis h under the optimal coding for hypothesis H.
- $-log_2 P(D \mid h)$ is the description length of the training data D given hypothesis h, using the optimal coding.

Thus, h_{MAP} is the hypothesis that minimizes the sum of the description length of the hypothesis plus the description length of the data given the hypothesis, given the optimal encoding.

MDL Principle: The best hypothesis is the one that minimizes the length of the hypothesis plus the length of the data given the hypothesis.

$$h_{MDL} = \text{argmin}_{h \in H} L_{C1}(h) + L_{C2}(D \mid h)$$

Using MDL to Avoid Overfitting in Decision Trees

- Given a decision tree h that classfies a fixed sequence of training instances D.
- How to encode the tree h?
- How to encode the examples? If the sequence of examples is fixed, then only the classification of the examples needs to be transmitted.
- If the tree has 0 error, then $L_{C1}(D \mid h) = 0$.
- Otherwise, for each misclassified example, we need $log_2 m + log_2 k$ bits (assuming m examples and k classes).
- MDL provides a way of trading off a smaller tree which makes more error on the training set, vs. a larger tree that has 0 error.

Bayes Optimal Classifier

Question: What is the most probable *classification* given the training data? Is it possible to do better than the MAP hypothesis?

Bayes Optimal Classification:

$$h_{BO} \equiv \text{argmax}_{v_j \in V} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D)$$

where V is the set of possible classifications of the example.

It is called *Bayes optimal* because *no* other classification method using the same hypothesis space and prior knowledge can outperform it, on the average.

Example of Bayes Optimal Classification

Let there be 5 hypotheses h_1 through h_5.

$P(h_i \mid D)$	$P(F \mid h_i)$	$P(L \mid h_i)$	$P(R \mid h_i)$
0.4	1	0	0
0.2	0	1	0
0.1	0	0	1
0.1	0	1	0
0.2	0	1	0

Then, the MAP hypothesis suggests the robot should go forward (F).

What does the Bayes optimal procedure suggest?

Example of Bayes Optimal Classification

$$\sum_{h_i \in H} P(F \mid h_i) P(h_i \mid D) = 0.4$$

$$\sum_{h_i \in H} P(L \mid h_i) P(h_i \mid D) = 0.2 + 0.1 + 0.2 = 0.5$$

$$\sum_{h_i \in H} P(R \mid h_i) P(h_i \mid D) = 0.1$$

Thus, Bayes optimal recommends the robot turn left.

Naive Bayes Classifier

Consider a classification problem, where each instance is described by n attributes $a_1, \ldots, a_n$. The possible classes are $v_i \in V$. Using the MAP estimator, we get

$$v_{MAP} \equiv \text{argmax}_{v_j \in V} P(v_j \mid a_1, \ldots, a_n)$$

Rewriting using the Bayes rule, we get

$$\begin{aligned} v_{MAP} &= \text{argmax}_{v_j \in V} \frac{P(a_1, \ldots, a_n \mid v_j) P(v_j)}{P(a_1, \ldots, a_n)} \\ &= \text{argmax}_{v_j \in V} P(a_1, \ldots, a_n \mid v_j) P(v_j) \end{aligned}$$

Naive Bayes Classifier

Suppose $n = 10$, and each attribute a_i can take on 5 values. Let the number of classes be 10. Then, how many conditional probabilities do we need to estimate from the training data?

$$5^{10} * 10!!!$$

Now, suppose that we assume the attribute values to be conditionally independent, given the class. That is,

$$P(a_1, \ldots, a_n \mid v_j) = \Pi_i P(a_i \mid v_j)$$

Now, how many probabilities do we need to estimate?

$$(5)(10)(10) = 500!$$

Naive Bayes Classifier Procedure

Given a set of labeled training examples, where each example is described by n attribute values a_i.

- Estimate the apriori class probabilities $P(v_j)$ as the fraction of the training set that has label v_j.
- Estimate the conditional attribute probabilities $P(a_i \mid v_j)$ as the proportion of instances belonging to class v_j that have value a_i for the i^{th} attribute.
- Given a new unlabeled training instance, compute its classification as

$$v_{NB} \equiv \text{argmax}_{v_j \in V} P(v_j) \Pi_i P(a_i \mid v_j)$$

Note that there is no explicit search through the space of hypotheses in NB, as there is in Bayes optimal classifiers.

Example: Document Classification

The problem of document classification is to assign a given document (text) into one of a set of categories.

Examples:

- Newsgroup articles I find interesting to read.
- Web pages that pertain to a particular topic.
- Library books that discuss Markov decision processes.
- Newspaper articles on European currency unification (ECU)

What are the possible attributes for a text document?

Example: Document Classification

"The countdown resumed Tuesday for the launch of NASA's controversial Cassini probe to Saturn after engineers fixed a technical problem at the launch pad. NASA has rescheduled the beginning of the $3.4 billion mission to explore the ringed planet for 4:43 a.m. EDT Wednesday. Cassini's Titan 4B rocket was supposed to have left Monday, but was delayed by a problem with a battery testing device on the launch pad, minor computer glitches and high upper-level winds that posed a safety threat in the event of an explosion during launch."

Attributes: $a_1 =$ "the", $a_2 =$ "countdown", ..., $a_{93} =$ launch.

If we apply Naive Bayes to classify this article (say into one of four categories: arts, science, music, sports), then

$$v_{NB} \equiv \text{argmax}_{v_j \in A, Sc, M, Sp} P(v_j) P(a_1 = \text{“the”} \mid v_j)$$

$$P(a_2 = \text{“countdown”} \mid v_j) \ldots P(a_{93} = \text{“launch”} \mid v_j)$$

Example: Document Classification

How many probabilities do we need to infer for document classification? Assume a maximum document length of 100 words, and 5 possible categories. Assuming 50000 words in English, we get

$$(50000)(100)(5) = 25 \text{ million!!!}$$

Even with the independence assumption, Naive Bayes seems hopeless for this task. We can, however, make an additional assumption that the conditional probability of a word given a classification is independent of the word location. That is,

$$P(a_i = \text{``NASA''} \mid v_j) = P(a_k = \text{``NASA''} \mid v_j)$$

So, for our example, this means that

$$\begin{aligned} v_{NB} \equiv \text{argmax}_{v_j \in A, Sc, M, Sp} P(v_j) P(\text{``the''} \mid v_j) \\ P(\text{``countdown''} \mid v_j) \ldots P(\text{``launch''} \mid v_j) \end{aligned}$$

Number of probabilities needed now is down to $(50000)(5) = 250,000$.

Estimating Probabilities from Small Sample Sizes

When the number of instances is small, frequency estimates of probabilities can be very wrong! For example, if the word "NASA" appears 10 times in 100 science documents, can we assume that $P(\text{``NASA''} \mid \text{``science''}) = 0.1$? Alternatively, if it never appears in any document in our collection, is the probability of seeing it in a collection of science documents 0?

m-estimates of probability:

$$\frac{n_c + mp}{n + m}$$

where n_c is the number of occurences of the word c in n documents, m is the equivalent sample size, and p is the prior probability estimate.

A common approach in document classification is to assume maximum likelihood (uniform priors), which would mean $m =$ vocabulary size.

$$\frac{n_c + 1}{n + \ |\text{vocabulary}|}$$

LEARN_NAIVE_BAYES_TEXT Procedure

1. Let vocabulary = set of all distinct words that occur in the document list (except for most frequent words, and least frequent words).
2. For each class v_j do
   - $P(v_j) = \frac{\text{docs}_j}{\text{examples}}$
   - For each word w_k do $P(w_k \mid v_j) = \frac{n_k+1}{n+|\text{Vocabulary}|}$

Given a new document, classify it into the category maximizing

$$v_{NB} \equiv \text{argmax}_{v_j \in V} P(v_j)\Pi_i P(a_i \mid v_j)$$

Test of NB_Text using Libbow/Rainbow – McCallum

1000 articles from rec.autos, alt.atheism, and sci.med.

```
Class ‘rec.autos’
  Gathering stats... files : unique-words ::     1000 :  13869
Class ‘sci.med’
  Gathering stats... files : unique-words ::     1000 :  27028
Class ‘alt.atheism’
  Gathering stats... files : unique-words ::     1000 :  34230

Correct: 986 out of 990 (99.60 percent accuracy)


 - Confusion details, row is actual, column is predicted
     classname   0   1   2  :total
 0   rec.autos 330   .   2  :332  99.40\%
 1 alt.atheism   1 326   1  :328  99.39\%
 2     sci.med   .   . 330  :330 100.00\%


Average_percentage_accuracy_and_stderr 99.60 0.00
```

Comparison with TFIDF

1000 articles from rec.autos, alt.atheism, and sci.med.

```
Class ‘rec.autos’
  Gathering stats... files : unique-words ::     1000 :  13869
Class ‘sci.med’
  Gathering stats... files : unique-words ::     1000 :  27028
Class ‘alt.atheism’
  Gathering stats... files : unique-words ::     1000 :  34230

Correct: 929 out of 990 (93.84 percent accuracy)


 - Confusion details, row is actual, column is predicted
     classname   0   1   2  :total
 0   rec.autos 299   8  19  :326  91.72%
 1 alt.atheism   5 314  13  :332  94.58%
 2     sci.med   6  10 316  :332  95.18%

Average_percentage_accuracy_and_stderr 93.84 0.00
```