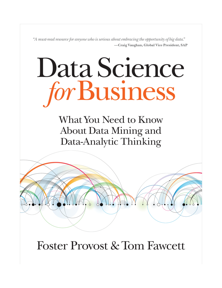
**Topic 7: Data Mining and Machine Learning. Naïve Bayes and Text Mining**



* + 1. **Evidence and Probabilities** Chapter 9
    2. **Define independent events**

Definition: Two events, A and B, are independent if the fact that A occurs does not affect the probability of B occurring.

* + 1. **Calculate the joint probability of two events.**

Joint probability is a statistical measure that calculates the likelihood of two events occurring together and at the same point in time.

The calculation of the joint probability is sometimes called the fundamental rule of probability or the “product rule” of probability.

P(A)=0.1666  
P(B)=0.1666

P(A,B)=0.1666 [x](https://investinganswers.com/dictionary/x/x) 0.1666)=0.02777

* + 1. **Recognize and apply joint probability using conditional probability**

A picture containing indoor, table

Description automatically generated

﻿

* + 1. **Calculate joint probability for independent and dependent events.**



Note: Bill Murray had six boys. Joint probability of these independent events =(½)6

Independent = MULTIPLY

Dependent = ADD

* + 1. **Explain the Bayes’ Rule (aka theorem) with the help of an example.**

A person posing for the camera

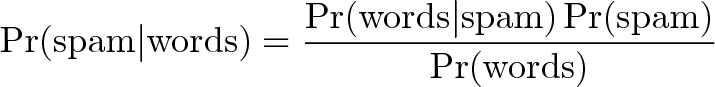
Description automatically generated Minister Thomas Bayes. Note: A basso.

**Describes the probability of an event, based on prior knowledge of conditions that might be related to the event.**

Spam example:

The **event**in this case is that the message is spam. The **test**for spam is that the message contains some flagged words (like “Viagra” or “you have won”).

Here’s the equation set up (from Wikipedia), read as “The probability a message is spam given that it contains certain flagged words”:



* + 1. **Define posterior probability, prior, likelihood, and conditional independence.**

﻿

A picture containing drawing, table

Description automatically generated

﻿Bayes’ Rule says that we can compute the probability of our hypothesis H given some evidence E.

Hypothesis and Evidence example:

H = patient has measles.

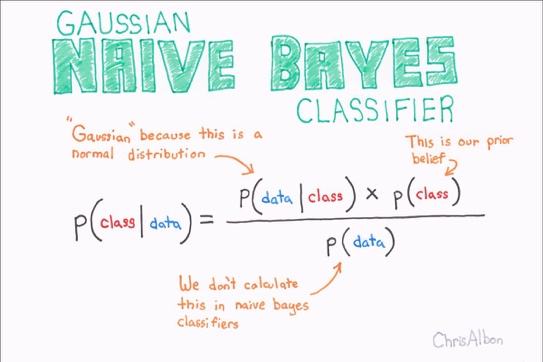
E = patient has red spots.

﻿LIKELIHOOD: p(E|H) is the probability that one has red SPOTS given that one has MEASLES. An expert in infectious diseases may well know this or be able to estimate it relatively accurately.

﻿PRIOR PROBABILITY: p(H) is simply the probability that someone has MEASLES, without considering any evidence; that’s just the prevalence of measles in the population. The degree to which we believe a given model accurately describes the given situation given the available data and all of our prior information.

﻿PROBABILITY OF EVIDENCE: p(E) is the probability of the evidence: what’s the probability that someone has red SPOTS — simply the prevalence of red spots in the population, which does not require complicated reasoning about the different underlying causes, just observation and counting.

POSTERIOR PROBABILITY: p(H|E) is the probability that one has MEASLES given that one has red SPOTS. It is the revised or updated probability of an event occurring after taking into consideration new information.



In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature



* + 1. **Explain the Naïve Bayes classifier.**

﻿A close up of a logo

Description automatically generated

The naïve Bayes classifier classifies a new example by estimating the probability that the example belongs to each class and reports the class with highest probability.

* + 1. **Explain why we do NOT need to calculate the denominator of the Bayes’ rule for naïve Bayes classifier.**

﻿A close up of a logo

Description automatically generated

First, if we are interested in classification, what we mainly care about is: of the different possible classes c, for which one is p (C| E) the greatest? In this case, E is the same for all, and we can just look to see which numerator is larger.

﻿Second, in cases where we would like the actual probability estimates, we still can get around computing p(E) in the denominator. This is because the classes often are mutually exclusive and exhaustive, meaning that every instance will belong to one and only one class.

* + 1. **List the advantages and disadvantages of the Naïve Bayes classifier.**

ADVANTAGES:

* + Very simple yet takes into account all features.
  + EASY to implement and fast (less computing time). Quick and dirty.
  + Incremental learner. Great for a user marking what is junk mail.
  + Need less training data.
  + Highly scalable. It scales linearly with the number of predictors and data points.
  + Handles continuous and discrete data.
  + Not sensitive to irrelevant features.
  + Works great in practice even when independence assumptions do not hold.

DISADVANTAGES:

* + Naive Bayes implicitly assumes that all the attributes are mutually independent.
  + Independence assumption not good for ranking when they do not hold.
  + Zero frequency problem. Have to use Laplace for smoothing. Sunrise problem.
  1. **Define generative model, lift, and Naïve-Naïve Bayes.**

Generative Model:

In statistical classification, including machine learning, two main approaches are called the generative approach and the discriminative approach.

Generative models can generate new data instances.

Discriminative models discriminate between different kinds of data instances.

Generative models tackle a more difficult task than analogous discriminative models. Generative models have to model *more*.



A generative model for images might capture correlations like "things that look like boats are probably going to appear near things that look like water" and "eyes are unlikely to appear on foreheads." These are very complicated distributions.

In contrast, a discriminative model might learn the difference between "sailboat" or "not sailboat" by just looking for a few tell-tale patterns. It could ignore many of the correlations that the generative model must get right.

Discriminative models try to draw boundaries in the data space, while generative models try to model how data is placed throughout the space. For example, the following diagram shows discriminative and generative models of handwritten digits:

Two graphs, one labelled 'Discriminative Model'
          and the other labelled 'Generative Model'. Both graphs show
          the same four datapoints. Each point is labeled with the image
          of the handwritten digit that it represents. In the discriminative
          graph there's a dotted line separating two data points from the
          remaining two. The region above the dotted line is labelled 'y=0' and
          the region below the line is labelled 'y=1'. In the generative graph
          two dotted-line circles are drawn around the two pairs of points. The
          top circle is labelled 'y=0' and the bottom circle is labelled 'y=1

Lift:

﻿ Lift measures how much more prevalent the positive class is in the selected subpopulation

over the prevalence in the population as a whole.

Portion of the sample and its performance. 3x,5x,10x etc.

A close up of a map

Description automatically generated

Naïve-Naïve Bayes:

﻿ Assume full feature independence, rather than the weaker assumption of conditional independence used for Naive Bayes. Let’s call this Naive-Naive Bayes, since it’s making stronger simplifying assumptions about the world.

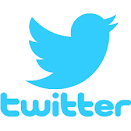
It is related to Lift.

A picture containing table

Description automatically generated

**7.1.2 Broad Issues involved in mining text chap. 10.**

1. **Explain why text is “dirty” which makes mining text is difficult**



Noisy text is text with differences between the surface form of a coded representation of the text and the intended, correct, or original text.

The noise may be due to typographic errors or colloquialisms always present in natural language and usually lowers the data quality in a way that makes the text less accessible to automated processing by computers, including natural language processing.

**7.1.3 Text representation** Provost

1. **Understand the meaning of “terms” when used in the field of Information Retrieval**

Information Retrieval (IR) is the science of searching for *information in a document*, searching for documents themselves, and also searching for the [metadata](https://en.wikipedia.org/wiki/Metadata) that describes data, and for databases of texts, images or sounds. Think of ‘terms’ as words or tokens.

1. **Describe the bag-of-words approach including the following steps: TF, IDF, TF-IDF.**

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR).

In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity.

1. Measuring term frequency (TF)

Term Frequency (TF), which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length.

1. Measuring sparseness (IDF)

The Inverse Document Frequency is a measure of how much information the word provides, i.e., if it's COMMON or RARE across all documents.

1. Combining them (TF-IDF)

TF-IDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

The TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general.

So, the word “the” has much less of an impact.

1. **Apply appropriate methods to create a TF-IDF representation of a query.**

Methods:

1. Basic stemming is applied
2. Stop words are removed
3. TFIDF
4. Use similarity functions like Cosine Similarity
5. **Express entropy in terms of the IDF measure.**

IDF and entropy are somewhat similar.

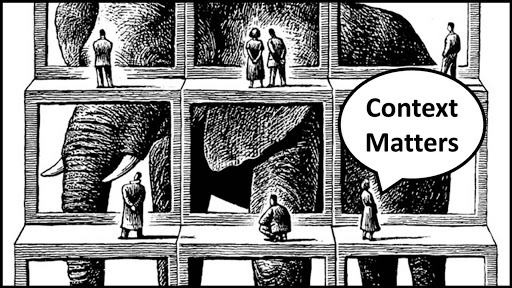
When no relevance information is available *(maximum entropy)*, the classical IDF corresponds to the measure of discriminativeness.

Maximum Entropy = no relevant information available

* + 1. **Additional text representation** Provost

1. **Describe N-Gram sequences**

This is a tool to understand CONTEXT. What comes before and after a word. It is like a context window so the machine can understand the meaning.



“We need to book our tickets soon.” VS. “We need to read this book soon.”

Bi-grams are usually the best. Before and after word to capture meaning.

Capture sarcasm. Pairs of words. “That is funny… not.”

1. **Describe Named Entity Abstraction (NEA)**

Also known as **chunking**. Putting text into CATEGORIES

 Note: “Sloth and Chunk [goonies]”

Example:

The Oakland Raiders are a football team and not a bunch of aggressive protestors.

Example:

Jim bought 300 shares of Acme Corp. in 2006.

And producing an annotated block of text that highlights the names of entities:

[Jim]Person bought 300 shares of [Acme Corp.]Organization in [2006]Time.

1. **Describe Topic Models**

Used in NLP and machine learning to figure out the topics in a corpus.

﻿The main idea of a topic LAYER is first to model the set of topics in a corpus separately.

A close up of a map

Description automatically generated

﻿One advantage is that in a search engine, for example, a query can use terms that do not exactly match the specific words of a document; if they map to the correct topic(s), the document will still be considered relevant to the search.

* + 1. **Mining news stories to predict stock price movement (Ch. 10)** Provost

1. **Describe how a certain task, such as recommending a news story that is likely to result in a significant change in the stock’s price, must be formulated into a problem with SIMPLYFYING ASSUMPTIONS.**



We want to predict stock price changes based on financial news.

SIMPLIFIED ASSUMPTIONS:

1. It is difficult to predict the effect of news far in advance. ﻿Therefore, we’ll try to predict what effect a news story will have on stock price the same day.
2. It is difficult to predict exactly what the stock price will be. Simplify it to direction. Simplify again to just change or no change.
3. It is difficult to predict small changes in stock price, so instead we’ll predict relatively large changes. Just how large is complicated. Simplify it to change (positive) and stable (negative).
4. We will assume that only news stories mentioning a specific stock will affect that stock’s price. Not true but it simplifies the task.
5. **Describe required considerations for data preprocessing**

Data preprocessing includes cleaning, instance selection, normalization, transformation, feature extraction and selection, etc.

Pre-process.

Stock news requires news all the time. Not just open and close.

Bag of words with a TF-IDF.

Remove stop words, stemmed, normalized, n-grams.

Each story is tagged as change or no change.

1. **Identify and discuss appropriate methods for analyzing the results**

You could do cost benefit analysis. Better to do ROC curve and Lift curve.

A close up of a map

Description automatically generated

A close up of a map

Description automatically generated

* + 1. **Classification** Article Jurafsky

**A close up of a logo

Description automatically generated**

1. **Describe typical applications of classifying text**

* Sentiment analysis
* Topic Labeling
* Language detection
* Intent detection

1. **Describe tasks involved in classifying text**

Rule-based Systems

Machine Learning Based Systems

1. **Compare alternative methods of classification**

Bootstrapping.

Bagging. Which is aggregate bootstrapping.

Boosting.

* + 1. **Math behind Naive Bayes classifiers**

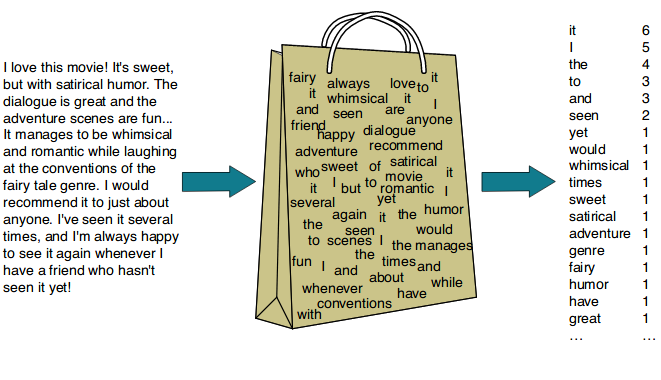
1. **Explain why in the context of classifying a document the denominator can be dropped from Bayes Rule**

Rigid independence of the variables.

Therefore, it is more proper to call Simple Bayes or Independence Bayes.

As described in Section Posterior Probabilities the posterior probability is the product of the class-conditional probability and the prior probability; the evidence term in the denominator can be dropped since it is **constant** for both classes.

1. **Explain Bag-of-words and naïve Bayes classifier assumptions**



It is called a “*bag*” of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document.

Naive Bayes is so called because the independence assumptions we have just made are indeed very naive for a model of natural language. The conditional independence assumption states that features are independent of each other given the class. This is hardly ever true for terms in documents.

1. **Explain why naïve Bayes calculations are done in log space so that predicted class is a linear function of input features**

Naïve Bayes calculations are done in a log space to avoid overflow in increase of speed.

A close up of a map

Description automatically generated Note: Log Curve

**7.2.3 Training the Naïve Bayes classifiers**

**A. Explain why text categorization often uses Laplace smoothing.**

**A screenshot of a cell phone

Description automatically generated**

In sports, the word **‘close’** in the corpus doesn’t appear as a final score**.** You either win or you lose. So, you can’t have a probability of zero.

By using something called Laplace smoothing: we add 1 to every count so it’s never zero. To balance this, we add the number of possible words to the divisor, so the division will never be greater than 1.

Laplace smoothing: you add fake samples to make bring your probability closer to a uniform probability.

 Note: Fake samples

A solution would be Laplace smoothing, which is a technique for smoothing categorical data.

A small-sample correction, or pseudo-count, will be incorporated in every probability estimate.

This is a way of regularizing Naive Bayes, and when the pseudo-count is zero, it is called Laplace smoothing

1. **Explain how to treat stopwords and unknown words during training.**



In computing, stop words are words which are filtered out before processing of natural language data.

Stop words are generally the most common words in a language; there is no single universal list of stop words used by all-natural language processing tools, and indeed not all tools even use such a list.

1. **Calculate the prior probabilities of two classes given a training set categorized into two categories.**

A picture containing knife

Description automatically generated

1. **Determine the class that a test sentence belongs to using the Naïve Bayes classifier**

Classify the sentence. Positive or Negative?

A screenshot of a cell phone

Description automatically generated

* + 1. **Optimizing for sentiment analysis**

1. **Explain how binary multinomial Naïve Bayes (Binary NB) differs from Naïve Bayes**

* NB stops the word counts in each document at 1.
* Binary Multinomial Bayes does a full count of positive and negative.
* Notice sentiment signs on left margin.

Example:

A screenshot of a cell phone

Description automatically generated

1. **Explain why binary multinomial Naïve Bayes (binary NB) might improve results relative to the standard Naïve Bayes approach**

* Gets rid of duplicates.
* Understands double negatives.

A screenshot of a cell phone

Description automatically generated

1. **Describe two other methods (besides binary NB) that can improve the results of sentiment analysis.**

In some situations, we might have insufficient training data. In such cases, we can derive

positive and negative word features from sentiment lexicons, lists of words that are pre-

annotated with positive or negative sentiment.

Two good ones are:

* + - 1. The General Inquirer
      2. LIWC ‘Luke’
    1. **Evaluation of sentiment analysis results**

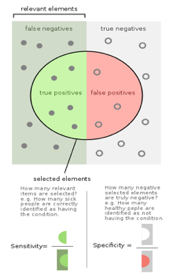
1. **Calculate precision and recall statistics given system output and gold standard label results.**

Gold labels are what humans have defined.

A screenshot of a cell phone

Description automatically generatedNote: Accuracy at an angle

A picture containing text

Description automatically generated

1. **Describe the F-score measure and various methods of weighting precision and recall**

In statistical analysis of binary classification, the F₁ score is a measure of a test's accuracy.

It considers both the precision and the recall of the test to compute the score.

Harmonic mean is used because it is a conservative metric. The harmonic value is closer to the minimum of two values than the arithmetic mean. It weighs the lower of the two values more heavily.

1. **Compare MACRO-averaging and MICRO-averaging approaches to evaluating the categorization performance of multiple classes**

* MICRO averaging is dominated by the more frequent class.
* MACRO averaging reflects the statistics of the smaller classes. Opposite.

A screenshot of a cell phone

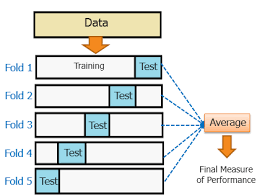
Description automatically generated

1. **Compare 10-fold cross validation VS. bootstrap tests**



Cross-validation uses ALL of your data points, whereas bootstrapping, which resamples your data randomly, may not hit all the points.

You can bootstrap as long as you want, meaning a larger resample, which should help with smaller samples.

****

Cross validation is a procedure for validating a model's performance, and it is done by splitting the training data into k parts.

k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data.

The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data.

The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation.

The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.