# Naive Bayes Classification

Emmanuel Rachelson

Suppose a feature space $\mathcal{X}$ and a class space $\mathcal{Y}$.
$X$ denotes the input random variable and $Y$ the output random variable.

**Example** (Vehicle classification)**.** *The automated e-mail system of an insurance company directs incoming requests of customers towards the most adequate employees, based on employee speciality: motorcycle, private cars, professionnal trucks. When the customer fills in a request, he indicates information about his vehicle, such as weight, color, number of wheels, energy source, horse power and the presence of a multimedia system. Based on this information, one needs to automatically classify the vehicles into the correct categories in order to send the requests to the right person.*
*So* $\mathcal{Y} = \{Mo, Ca, Tr\}$ *(*$Mo$ *for motorcycles,* $Ca$ *for cars,* $Tr$ *for trucks) and* $X$ *is composed of 6 variables:* $W$ *(weight, continuous),* $C$ *(color, categorical),* $N$ *(number of wheels, discrete ordered),* $E$ *(energy source, categorical),* $H$ *(horse power, continuous) and* $M$ *(multimedia, categorical binary).*
*The company has some evidence data* $\mathcal{D} = \{(x_i, y_i)\}$ *and wishes to build the function* $f$ *that will assign the correct label to incoming requests.*

## Bayes classifier

**Example** (Most probable vehicle class)**.** *To correctly classify an incoming message, the selection function should estimate the probability that the request is for* $Y = Mo$, $Y = Ca$ *and* $Y = Tr$ *given the value taken by* $X = (W, C, N, E, H, M)$*. That is, estimate the conditional probability* $\mathbb{P}(Y|X)$ *for all values of* $y$ *and pick the value of* $y$ *that maximizes this probability.*

**Definition** (Bayes classifier)**.** *Given an input* $x$*, the Bayes classifier assigns to* $x$ *the class that maximizes the conditional probability:*

$$f(x) = \underset{y\in\mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}(Y = y|X = x)$$

**Theorem** (Bayes theorem)**.** *Given two events* $A$ *and* $B$*,*

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

## Prior, posterior, likelihood, evidence

The Bayes classifier chooses the most probable class $y$ given input $x$. But according to Bayes theorem, this probability estimate is:

$$\mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(Y = y)\mathbb{P}(X = x|Y = y)}{\mathbb{P}(X = x)}$$

Estimating $\mathbb{P}(Y = y|X = x)$ boils down to estimating the probability of $y$ after we observe $x$, thus it is called the *posterior* probability estimate. Following that line, $\mathbb{P}(Y = y)$ is the uninformed probability estimate for class $y$, so it is called the *prior* probability estimate (prior to the observation of $x$). $\mathbb{P}(X = x|Y = y)$ is the *likelihood* of observing an individual (input) $x$ from class $y$. And finally $\mathbb{P}(X = x)$ is the *evidence*, indicating how likely it is to observe $x$ in general. So:

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

## The naive assumption

**Example.** *The maximum likelihood estimator for the prior $\mathbb{P}(Y = y)$ is $\frac{N_y}{N}$ where $N_y$ is the number of observations of class $y$ and $N$ is the total number of observations in $\mathcal{D}$. Hence, it is easy to store this prior in a table (one entry per value of $Y$).*
*Estimating the evidence $\mathbb{P}(X = x)$ has no utility when one tries to discriminate between different values of $y$. So the burden of estimating the evidence distribution can be avoided.*
*Finally, given a new request with parameters $X = x$, one needs to estimate the likelihood of this request for each vehicle category $Y = y$. This likelihood is:*

$$\mathbb{P}(X = x|Y = y) = \mathbb{P}(W = w, C = c, N = n, E = e, H = h, M = m|Y = y)$$

*for simplicity, we write this probability $\mathbb{P}(W, C, N, E, H, M|Y)$. Since for any two events $A$ and $B$, $\mathbb{P}(A, B) = \mathbb{P}(A|B)\mathbb{P}(B)$, we can decompose this probability as:*

$$\mathbb{P}(X = x|Y = y) = \mathbb{P}(W, C, N, E, H|Y, M) \cdot \mathbb{P}(M|Y)$$

*And so on until:*

$$\begin{aligned}\mathbb{P}(X = x|Y = y) =& \mathbb{P}(W|Y, M, H, E, N, C)\\ &\cdot \mathbb{P}(C|Y, M, H, E, N)\\ &\cdot \mathbb{P}(N|Y, M, H, E)\\ &\cdot \mathbb{P}(E|Y, M, H)\\ &\cdot \mathbb{P}(H|Y, M)\\ &\cdot \mathbb{P}(M|Y)\end{aligned}$$

*The probability of presence of a multimedia system in each category $\mathbb{P}(M|Y)$ is quite easy to assess from the data $\mathcal{D}$. To store the maximum likelihood estimator for $\mathbb{P}(M|Y)$, one just needs to store frequencies of presence of multimedia systems for each value of $Y$.*

*The probability $\mathbb{P}(H|Y,M)$ of a given horse-power, given the vehicule category and the presence of a multimedia system is slightly trickier. Since $H$ is a continuous variable, one needs to choose a family of distributions that can describe the distribution of engine power for, for instance, trucks equipped with a multimedia system. Once this family of distributions is chosen (here a Gaussian distribution might make sense), then it boils down to estimating its parameters for each combination of vehicle category ($Y$) and $M$.*

*This raises a question: it is reasonable to assume that the engine power is uncorrelated with the presence of a multimedia system, and thus $\mathbb{P}(H|Y,M) = \mathbb{P}(H|Y)$. Such an assumption simplifies the estimation problem since it divides by two the number of combinations of $(Y,M)$.*

*The next step is estimating $\mathbb{P}(E|Y,M,H)$, that is the likelihood of observing a vehicle using energy source $E$, given its category $Y$, its multimedia system $M$ and its engine power $H$. Since $H$ is continuous, this distribution over the categorical variable $E$ has a parametric form that depends continuously on the value taken by $H$. And, contrarily to the previous case, it is no more reasonable to assume that energy source $E$ and engine power $H$ are uncorrelated so it would be abusive and very naive to assume that $\mathbb{P}(E|Y,M,H) = \mathbb{P}(E|Y)$.*

*The same problem arises for the univariate distributions $\mathbb{P}(N|Y,M,H,E)$, $\mathbb{P}(C|Y,M,H,E,N)$ and $\mathbb{P}(W|Y,M,H,E,N,C)$. A few independence assumptions might be reasonable but the complexity of estimating these distributions quickly becomes prohibitive.*

*Question: what if one makes this rather naive assumption anyway?*

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}$, constructing a Bayes classifier implies constructing a probability estimator of $\mathbb{P}(Y = y|X = x)$. Bayes theorem shows that it requires estimating the prior $\mathbb{P}(Y = y)$ and likelihood $\mathbb{P}(X = x|Y = y)$ distributions. To discriminate between two classes, estimating the evidence is not necessary since $\mathbb{P}(X = x)$ does not depend on $y$.

Estimating $\mathbb{P}(Y = y)$ is relatively straightforward. Estimating $\mathbb{P}(X = x|Y = y)$ on the other hand requires estimating a multivariate distribution, which might be difficult if the number of features in $X$ is large. The *naive conditional independence assumption* simplifies this estimate.

**Definition** (Naive conditional independence assumption)**.** *This assumption states that any two variables $X_i$ and $X_j$ in $X$ are conditionaly independent given $Y$, that is:*

$$\forall i \neq j, \mathbb{P}(X_i|Y, X_j) = \mathbb{P}(X_i|Y)$$

Under the naive conditional independence assumption, the posterior estimate simplifies to:

$$\mathbb{P}(Y|X_1,\dots,X_n) = \frac{1}{Z} \times \mathbb{P}(Y) \times \prod_{i=1}^{n} \mathbb{P}(X_i|Y)$$

Where $Z$ is the normalizing factor corresponding to the evidence $\mathbb{P}(X)$ Each distribution $\mathbb{P}(X_i|Y)$ is a univariate distribution that is estimated using the data at hand $\mathcal{D}$.

**Definition** (Training a Naive Bayes Classifier)**.** *Training of a Naive Bayes classifier consists in estimating the parameters of $\mathbb{P}(Y = y)$ and $\mathbb{P}(X_i|Y = y)$ for all possible values of $y$.*

**Definition** (Prediction using a Naive Bayes Classifier)**.** *For a new input $x$, predicting the associated class using a Naive Bayes Classifier requires computing*

$$\operatorname*{argmax}_{y} \mathbb{P}(Y = y) \times \prod_{i=1}^{n} \mathbb{P}(X_i = x_i|Y = y) \tag{1}$$

## Practical issues

**Log-likelihood instead of likelihood.** Since estimated probabilities are often small numbers, it is often preferable to express Equation 1 in logarithmic form to avoid numerical errors:

$$\operatorname*{argmax}_{y\in[1,k]} \left[ \log \mathbb{P}(Y = y) + \sum_{i=1}^{n} \log \mathbb{P}(X_i = x_i|Y = y) \right] \tag{2}$$

**Choosing parametric distributions.** Depending on the nature (continuous / categorical) of each variable $X_i$ and the knowledge of its meaning, one needs to make a informed choice for the parametric form of the $\mathbb{P}(X_i = x_i|Y = y)$ distribution.

Common choices for continuous variables: normal, log-normal, Gamma, Poisson distributions.

E.g.: choosing a normal distribution $\mathbb{P}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ implies estimating the $\mu$ and $\sigma$ parameters from the data.

Common choices for categorical variables: Multinomial/categorical distributions.

E.g.: choosing a categorical distribution $\mathbb{P}(x) = \prod_{i=1}^{k} p_i^{(x=i)}$, implies estimating the occurence probability $p_i$ of each category.

**Laplace / Lidstone smoothing.** Consider a categorical variable $X_i$ that can take $k$ values and suppose that a given value $d$ of $X_i$ has never been observed in the data. Does that mean that the estimated $\mathbb{P}(X_i = d|Y = y)$ should be zero? If so, then this zero value will simply pull the whole expression of Equation

1 down to zero. To circumvent this issue, one often uses *additive smoothing* (also known as Laplace smoothing or Lidstone smoothing) and writes:

$$\mathbb{P}(X_i = d|Y = y) = \frac{N_{dy} + 1}{N_y + k}$$

where $N_{dy}$ is the number of occurences of $X_i = d|Y = y$ while $N_y$ is the number of occurences of $Y = y$. This way, as the amount of i.i.d. data grows, this estimator converges from a uniform distribution to the true distribution of $X_i|Y$.

**Bayesian networks.** Bayesian networks are a graphical representation of conditional dependencies between variables, where variables are nodes and a directed arc between two variables indicates the conditional dependency of the pointed variable on the origin variable. The naive assumption states that all $X_i$ are conditionally independent. Hence the Bayesian networks corresponding to a Naive Bayes classifier has only arcs going from $X_i$ variables to $Y$. If, however, some knowledge about the interdependencies of the $X_i$ are known beforehand, then the representative Bayesian network will have arcs between the corresponding $X_i$. Then the decomposition *by variable* made possible by the naive assumption may be refined into a decomposition by *groups* of variables, instead of individual variables. This comes at the price of a more complex model to estimate, but such structural knowledge can be useful to attain better performance for Naive Bayes classifiers (although in general performance is satisfactory with the raw naive assumption).

## Optimality of Naive Bayes classifiers

The naive conditional independence assumption (also called "naive Bayes assumption" for brevity) is a very strong one which obviously does not hold in the vast majority of real-world situations. Nevertheless, naive Bayes classifiers perform surprisingly well in practice. The key to this performance lies in the fact that although the naive Bayes assumption makes the $\mathbb{P}(Y|X)$ estimator a very poor one, this one still affects a higher priority to the most probable class, thus yielding a correct answer to the argmax operator. A thorough analysis of this phenomenon can be found in: **The Optimality of Naive Bayes**, H. Zhang, *FLAIRS*, 2004.

## Document classification, the Bag-Of-Words model

One very common application of naive Bayes classifiers is document classification (e-mail spam filtering, sentiment analysis on social networks, technical documentation classification, customer appreciations, etc.). Naive Bayes classifiers for documents estimate the probability of a given document belonging to a certain class $Y$ of documents, based on the document's contents $X_i$.

A popular choice of $X_i$ variables for describing the document's contents is the bag-of-word model. The bag-of-words corresponding to a document is the count of occurence of each of its words, regardless of grammar or ordering. Then $X_i$ is the number of occurences of the $i$th word of the dictionnary in the current document. Generally, dictionaries used in this context ignore stop-words (most commmon words in a language, like "the" or "is"), non-words (like "55.3" or punctuation) and documents are pre-processed by performing stemming and lemmatization ("be", "being", "is", "are" all fall into the same word "be").

The combination of the bag-of-words model, Laplace smoothing and naive Bayes classifiers generally provides flexible and efficient document categorization capabilities.