Bayes Classifier and Naive Bayes

Cornell CS 4/5780

Spring 2024

previous next

(Lecture 9) (Lecture 10)

Our training consists of the set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ drawn from some unknown distribution $P(X, Y)$. Because all pairs are sampled i.i.d., we obtain

$$P(D) = P((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) = \Pi_{\alpha=1}^{n} P(\mathbf{x}_\alpha, y_\alpha).$$

If we do have enough data, we could estimate $P(X, Y)$ similar to the coin example in the previous lecture, where we imagine a **gigantic** die that has one side for each possible value of $(\mathbf{x}, y)$. We can estimate the probability that one specific side comes up through counting:

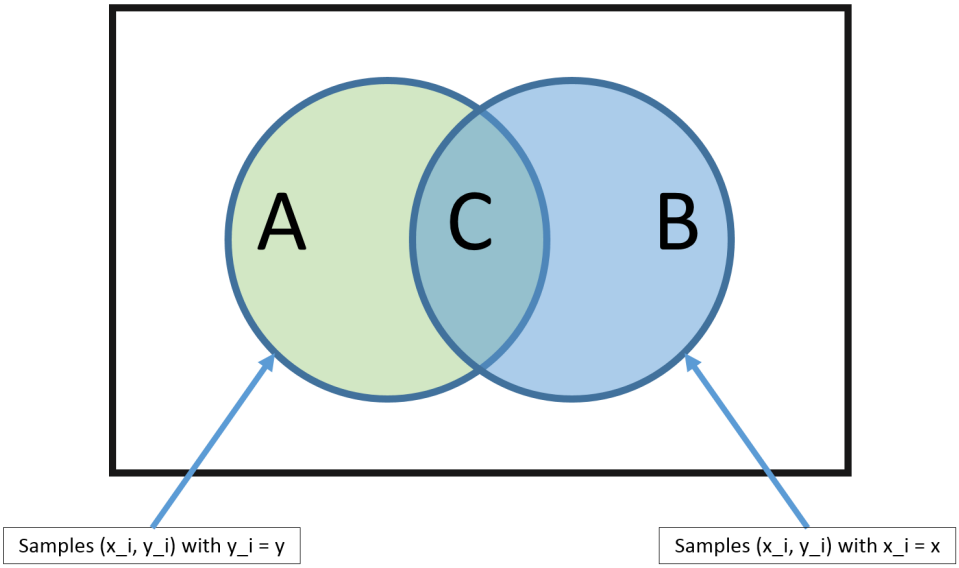
$$\hat{P}(\mathbf{x}, y) = \frac{\sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x} \wedge y_i = y)}{n},$$

where $I(\mathbf{x}_i = \mathbf{x} \wedge y_i = y) = 1$ if $\mathbf{x}_i = \mathbf{x}$ and $y_i = y$ and 0 otherwise.

Of course, if we are primarily interested in predicting the label y from the features $\mathbf{x}$, we may estimate $P(Y|X)$ directly instead of $P(X, Y)$. We can then use the Bayes Optimal Classifier for a specific $\hat{P}(y|\mathbf{x})$ to make predictions.

So how can we estimate $\hat{P}(y|\mathbf{x})$? Previously we have derived that $\hat{P}(y) = \frac{\sum_{i=1}^{n} I(y_i = y)}{n}$. Similarly, $\hat{P}(\mathbf{x}) = \frac{\sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x})}{n}$ and $\hat{P}(y, \mathbf{x}) = \frac{\sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x} \wedge y_i = y)}{n}$. We can put these two together

$$\hat{P}(y|\mathbf{x}) = \frac{\hat{P}(y, \mathbf{x})}{P(\mathbf{x})} = \frac{\sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x} \wedge y_i = y)}{\sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x})}$$



The Venn diagram illustrates that the MLE method estimates $\hat{P}(y|\mathbf{x})$ as

$$\hat{P}(y|\mathbf{x}) = \frac{|C|}{|B|}$$

Problem: But there is a big problem with this method. The MLE estimate is only good if there are many training vectors with the **same identical** features as $\mathbf{x}$. In **high dimensional spaces** (or with continuous $\mathbf{x}$), this never happens! So $|B| \to 0$ and $|C| \to 0$.

Naive Bayes

We can approach this dilemma with a simple trick, and an additional assumption. The trick part is to estimate $P(y)$ and $P(\mathbf{x}|y)$ instead, since, by Bayes rule,

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})}.$$

Recall from Estimating Probabilities from Data that estimating $P(y)$ and $P(\mathbf{x}|y)$ is called *generative learning*.
Estimating $P(y)$ is easy. For example, if Y takes on discrete binary values estimating $P(y)$ reduces to coin tossing. We simply need to count how many times we observe each outcome (in this case each class):

$$P(y = c) = \frac{\sum_{i=1}^{n} I(y_i = c)}{n} = \hat{\pi}_c$$

Estimating $P(\mathbf{x}|y)$, however, is not easy! The additional assumption that we make is the *Naive Bayes assumption*.

Naive Bayes Assumption:

$$P(\mathbf{x}|y) = \prod_{\alpha=1}^{d} P(x_\alpha|y), \text{ where } x_\alpha = x_\alpha \text{ is the value for feature } \alpha$$

i.e., feature values are **independent given the label!** This is a very **bold** assumption.

For example, a setting where the Naive Bayes classifier is often used is spam filtering. Here, the data is emails and the label is *spam* or *not-spam*. The Naive Bayes assumption implies that the words in an email are conditionally independent, given that you know that an email is spam or not. Clearly this is not true. Neither the words of spam or not-spam emails are drawn independently at random. However, the resulting classifiers can work well in practice even if this assumption is violated.

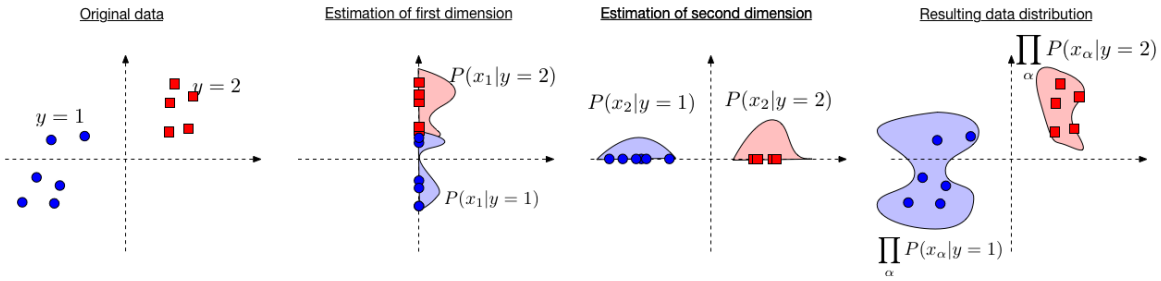


Illustration behind the Naive Bayes algorithm. We estimate $P(x_\alpha|y)$ independently in each dimension (middle two images) and then obtain an estimate of the full data distribution by assuming conditional independence $P(\mathbf{x}|y) = \prod_\alpha P(x_\alpha|y)$ (very right image).

So, for now, let's pretend the Naive Bayes assumption holds. Then the Bayes Classifier can be defined as

$$\begin{aligned}
h(\mathbf{x}) &= \operatorname*{argmax}_y P(y|\mathbf{x}) \\
&= \operatorname*{argmax}_y \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})} \\
&= \operatorname*{argmax}_y P(\mathbf{x}|y)P(y) && (P(\mathbf{x}) \text{ does not depend on } y) \\
&= \operatorname*{argmax}_y \prod_{\alpha=1}^{d} P(x_\alpha|y)P(y) && (\text{by the naive Bayes assumption}) \\
&= \operatorname*{argmax}_y \sum_{\alpha=1}^{d} \log(P(x_\alpha|y)) + \log(P(y)) && (\text{as log is a monotonic function})
\end{aligned}$$

Estimating $\log(P(x_\alpha|y))$ is easy as we only need to consider one dimension. And estimating $P(y)$ is not affected by the assumption.

Estimating $P(x_\alpha|y)$

Now that we know how we can use our assumption to make the estimation of $P(y|\mathbf{x})$ tractable. There are 3 notable cases in which we can use our naive Bayes classifier.

Case #1: Categorical features

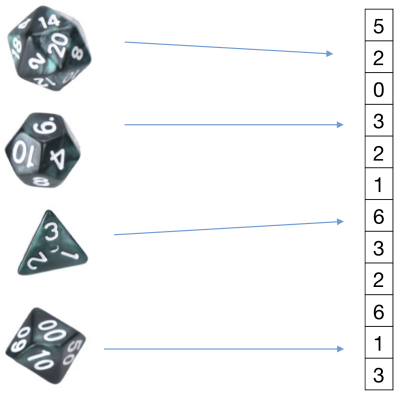


Illustration of categorical NB. For d dimensional data, there exist d independent dice for each class. Each feature has one die per class. We assume training samples were generated by rolling one die after another. The value in dimension i corresponds to the outcome that was rolled with the i^{th} die.

Features:

$$x_\alpha \in \{f_1, f_2, \cdots, f_{K_\alpha}\}$$

Each feature α falls into one of K_α categories. (Note that the case with binary features is just a specific case of this, where $K_\alpha = 2$.) An example of such a setting may be medical data where one feature could be *marital status* (single / married). Model $P(x_\alpha \mid y)$:

$$P(x_\alpha = j|y = c) = [\theta_{jc}]_\alpha \text{ and } \sum_{j=1}^{K_\alpha}[\theta_{jc}]_\alpha = 1$$

where $[\theta_{jc}]_\alpha$ is the probability of feature α having the value j, given that the label is c. And the constraint indicates that x_α must have one of the categories $\{1, \ldots, K_\alpha\}$. Parameter estimation:

$$[\hat{\theta}_{jc}]_\alpha = \frac{\sum_{i=1}^{n} I(y_i = c)I(x_{i\alpha} = j) + l}{\sum_{i=1}^{n} I(y_i = c) + lK_\alpha},$$

where $x_{i\alpha} = [\mathbf{x}_i]_\alpha$ and l is a smoothing parameter. By setting $l = 0$ we get an MLE estimator, and $l > 0$ leads to MAP. If we set $l = +1$ we get *Laplace smoothing*.

In words (without the l hallucinated samples) this means

$$\frac{\#\text{ of samples with label c that have feature } \alpha \text{ with value } j}{\#\text{ of samples with label } c}.$$

ssentially the categorical feature model associates a special coin with each feature and label. The generative model that we are assuming is that the data was generated by first choosing the label (e.g. *"healthy person"*). That label comes with a set of d "dice", for each dimension one. The generator picks each die, tosses it and fills in the feature value with the outcome of the coin toss. So if there are C possible labels and d dimensions we are estimating $d \times C$ "dice" from the data. However, per data point only d dice are tossed (one for each dimension). Die α (for any label) has K_α possible "sides". Of course this is not how the data is generated in reality - but it is a modeling assumption that we make. We then learn these models from the data and during test time see which model is more likely given the sample.

Prediction:

$$\underset{y}{\text{argmax}}\ P(y = c \mid \mathbf{x}) \propto \underset{y}{\text{argmax}}\ \hat{\pi}_c \prod_{\alpha=1}^{d} [\hat{\theta}_{jc}]_\alpha$$

Case #2: Multinomial features

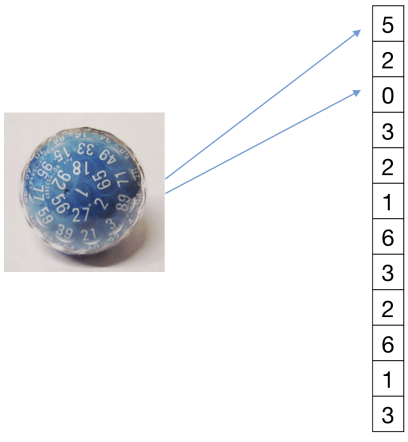


Illustration of multinomial NB. There are only as many dice as classes. Each die has d sides. The value of the i^{th} feature shows how many times this particular side was rolled.

If feature values don't represent categories (e.g. single/married) but counts we need to use a different model. E.g. in the text document categorization, feature value $x_\alpha = j$ means that in this particular document $\mathbf{x}$ the α^{th} word in my dictionary appears j times. Let us consider the example of spam filtering. Imagine the α^{th} word is indicative of being "spam". Then if $x_\alpha = 10$ means that this email is likely spam (as word α appears 10 times in it). And another email with $x'_\alpha = 20$ should be even more likely to be spam (as the spammy word appears twice as often). With categorical features this is not guaranteed. It could be that the training set does not contain any email that contain word α exactly 20 times. In this case you would simply get the hallucinated smoothing values for both spam and not-spam - and the signal is lost. We need a model that incorporates our knowledge that features are counts - this will help us during estimation (you don't have to see a training email with exactly the same number of word occurrences) and during inference/testing (as you will obtain these monotonicities that one might expect). The multinomial distribution does exactly that.

Features:

$$x_\alpha \in \{0, 1, 2, \ldots, m\} \text{ and } m = \sum_{\alpha=1}^{d} x_\alpha$$

Each feature α represents a count and m is the length of the sequence. An example of this could be the count of a specific word α in a document of length m and d is the size of the vocabulary. Model $P(\mathbf{x} \mid y)$: Use the multinomial distribution

$$P(\mathbf{x} \mid m, y = c) = \frac{m!}{x_1! \cdot x_2! \cdot \cdots \cdot x_d!} \prod_{\alpha=1}^{d} (\theta_{\alpha c})^{x_\alpha}$$

where $\theta_{\alpha c}$ is the probability of selecting x_α and $\sum_{\alpha=1}^{d} \theta_{\alpha c} = 1$. So, we can use this to generate a spam email, i.e., a document $\mathbf{x}$ of class $y = \text{spam}$ by picking m words independently at random from the vocabulary of d words using $P(\mathbf{x} \mid y = \text{spam})$. Parameter estimation:

$$\hat{\theta}_{\alpha c} = \frac{\sum_{i=1}^{n} I(y_i = c) x_{i\alpha} + l}{\sum_{i=1}^{n} I(y_i = c) m_i + l \cdot d}$$

where $m_i = \sum_{\beta=1}^{d} x_{i\beta}$ denotes the number of words in document i. The numerator sums up all counts for feature x_α and the denominator sums up all counts of all features across all data points. E.g.,

$$\frac{\#\text{ of times word } \alpha \text{ appears in all spam emails}}{\#\text{ of words in all spam emails combined}}.$$

Again, l is the smoothing parameter. Prediction:

$$\underset{c}{\text{argmax}}\ P(y=c \mid \mathbf{x}) \propto \underset{c}{\text{argmax}}\ \hat{\pi}_c \prod_{\alpha=1}^{d} \hat{\theta}_{\alpha c}^{x_\alpha}$$

Case #3: Continuous features (Gaussian Naive Bayes)

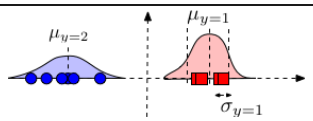


Illustration of Gaussian NB. Each class conditional feature distribution $P(x_\alpha|y)$ is assumed to originate from an independent Gaussian distribution with its own mean $\mu_{\alpha,y}$ and variance $\sigma^2_{\alpha,y}$.

Features:

$$x_\alpha \in \mathbb{R} \qquad \text{(each feature takes on a real value)}$$

Model $P(x_\alpha \mid y)$: Use Gaussian distribution

$$P(x_\alpha \mid y=c) = \mathcal{N}\left(\mu_{\alpha c}, \sigma^2_{\alpha c}\right) = \frac{1}{\sqrt{2\pi}\sigma_{\alpha c}} e^{-\frac{1}{2}\left(\frac{x_\alpha - \mu_{\alpha c}}{\sigma_{\alpha c}}\right)^2}$$

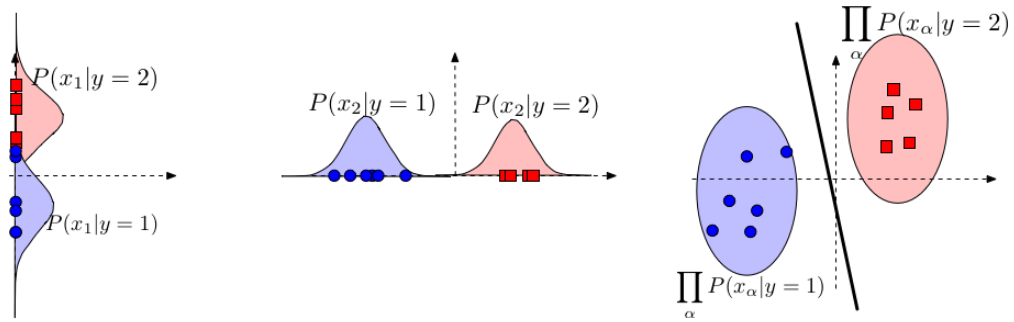
Note that the model specified above is based on our assumption about the data - that each feature α comes from a class-conditional Gaussian distribution. The full distribution $P(\mathbf{x}|y) \sim \mathcal{N}(\mu_y, \Sigma_y)$, where Σ_y is a diagonal covariance matrix with $[\Sigma_y]_{\alpha,\alpha} = \sigma^2_{\alpha,y}$.

Parameter estimation: As always, we estimate the parameters of the distributions for each dimension and class independently. Gaussian distributions only have two parameters, the mean and variance. The mean $\mu_{\alpha,y}$ is estimated by the average feature value of dimension α from all samples with label y. The (squared) standard deviation is simply the variance of this estimate.

$$\mu_{\alpha c} \leftarrow \frac{1}{n_c}\sum_{i=1}^{n} I(y_i = c) x_{i\alpha} \qquad \text{where } n_c = \sum_{i=1}^{n} I(y_i = c)$$

$$\sigma^2_{\alpha c} \leftarrow \frac{1}{n_c}\sum_{i=1}^{n} I(y_i = c)(x_{i\alpha} - \mu_{\alpha c})^2$$

Naive Bayes is a linear classifier



Naive Bayes leads to a linear decision boundary in many common cases. Illustrated here is the case where $P(x_\alpha|y)$ is Gaussian and where $\sigma_{\alpha,c}$ is identical for all c (but can differ across dimensions α). The boundary of the ellipsoids indicate regions of equal probabilities $P(\mathbf{x}|y)$. The red decision line indicates the decision boundary where $P(y=1|\mathbf{x}) = P(y=2|\mathbf{x})$.

1. Suppose that $y_i \in \{-1, +1\}$ and features are multinomial We can show that

$$P(y \mid \mathbf{x}) = \frac{1}{1 + e^{-y(\mathbf{w}^\top \mathbf{x} + b)}}.$$

As before, we define $P(x_\alpha|Y=+1) \propto \theta_{\alpha+}^{x_\alpha}$ and $P(Y=+1) = \pi_+$. Let us define two weight vectors $\mathbf{w}_+, \mathbf{w}_-$ and bias terms b^+, b^- as

$$w_\alpha^+ = \log[\theta_{\alpha+}]$$

$$b^+ = \log[P(Y=+1)].$$

First let us consider $\log\left[P(\mathbf{x}|Y=+1)\right]$ and remember that $P(x_\alpha \mid Y=+1) = \theta_{\alpha+}^{x_\alpha}$:

$$\begin{aligned}
\log\left[P(\mathbf{x}|Y=+1)\right] &= \log\left[\Pi_{\alpha=1}^d P(x_\alpha \mid Y=+1)\right] \\
&= \sum_{\alpha=1}^d x_\alpha \log[\theta_{\alpha+}] \\
&= \sum_{\alpha=1}^d x_\alpha w_\alpha^+ \\
&= \mathbf{x}^\top \mathbf{w}_+.
\end{aligned}$$

It follows that $P(\mathbf{x}|Y=+1) = e^{\mathbf{x}^\top \mathbf{w}_+}$ and $P(\mathbf{x}|Y=-1) = e^{\mathbf{x}^\top \mathbf{w}_-}$. Also, by definition $P(Y=+1) = e^{b^+}$ and $P(Y=-1) = e^{b^-}$. Let us define the differences between the two weight vectors and biases as: $\mathbf{w} = \mathbf{w}_- - \mathbf{w}_+$ and $b = b_- - b_+$. We can use Bayes Rule to derive:

$$
\begin{aligned}
P(Y=+1 \mid \mathbf{x}) &= \frac{P(\mathbf{x} \mid +1)P(Y=+1)}{P(\mathbf{x} \mid +1)P(Y=+1) + P(\mathbf{x} \mid -1)P(Y=-1)} \\
&= \frac{e^{\mathbf{x}^\top \mathbf{w}_+ + b_+}}{e^{\mathbf{x}^\top \mathbf{w}_+ + b_+} + e^{\mathbf{x}^\top \mathbf{w}_- + b_-}} \\
&= \frac{1}{1 + e^{-(\mathbf{x}^\top \mathbf{w} + b)}}
\end{aligned}
$$

Finally, because our labels $y \in \{+1, -1\}$ we can conveniently create one equation for both classes:

$$
P(Y=y \mid \mathbf{x}) = \frac{1}{1 + e^{-y(\mathbf{x}^\top \mathbf{w} + b)}}.
$$

2. The exact same equation can be drived for Gaussian Naive Bayes with constant variance (i.e. $\sigma_{\alpha 1} = \sigma_{\alpha -1}$ for all α), except that the vector $\mathbf{w}$ is here the difference of the means.