Naive Bayes Models

(Supervised Classification)

Lê Hồng Phương
Data Science Laboratory, VNU Hanoi
<*phuonglh@hus.edu.vn*>

March 22, 2020

Content

Classification, Regression

- Classification problem: classify each object $\mathbf{x} \in \mathcal{X}$ into one class $y \in \mathcal{Y} = \{1, 2, \ldots, K\}$.
- Vector representation: $\mathbf{x} = (x_1, x_2, \ldots, x_D)$.
  - We say that $\mathbf{x}$ has D *features* or *attributes*.
  - Each component x_j is called a *feature value* or an *attribute value* of $\mathbf{x}$.
  - x_j can be discrete or continuous ($x_j \in \mathbb{R}$).

Classification, Regression

- Classification problem: classify each object $\mathbf{x} \in \mathcal{X}$ into one class $y \in \mathcal{Y} = \{1, 2, \ldots, K\}$.
- Vector representation: $\mathbf{x} = (x_1, x_2, \ldots, x_D)$.
  - We say that $\mathbf{x}$ has D *features* or *attributes*.
  - Each component x_j is called a *feature value* or an *attribute value* of $\mathbf{x}$.
  - x_j can be discrete or continuous ($x_j \in \mathbb{R}$).
- Supervised learning: given a dataset of N samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$, we need to build a good model/hypothese $h(\cdot)$ to predict y given $\mathbf{x}$.
- Notation: $\mathbf{x}_i$ is the i-th sample of the dataset, x_{ij} is the j-th feature value of $\mathbf{x}_i$.
  - If $h(\mathbf{x})$ is discrete (finite), we have a **classification** problem.
  - If $h(\mathbf{x})$ is continuous (infinite) we have a **regression** problem.

Generative versus Discriminative Models

- A model is said generative if it can be used to generate the dataset given some hidden parameter $\theta \Rightarrow P(\mathbf{x}, y; \theta)$.
- Bayes's rule:

$$P(\mathbf{x}, y; \theta) = P(\mathbf{x} \,|y; \theta) P(y; \theta).$$

- So, in a generative model, we need to model the *likelihood distribution* $P(\mathbf{x} \,|y)$ and the *prior distribution* $P(y)$.

Generative versus Discriminative Models

- A model is said generative if it can be used to generate the dataset given some hidden parameter $\theta \Rightarrow P(\mathbf{x}, y; \theta)$.
- Bayes's rule:

$$P(\mathbf{x}, y; \theta) = P(\mathbf{x} \,|y; \theta) P(y; \theta).$$

- So, in a generative model, we need to model the *likelihood distribution* $P(\mathbf{x} \,|y)$ and the *prior distribution* $P(y)$.
- Next, we compute the *posterior distribution*:

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

- The classification rule for an object $\mathbf{x}$ is:

$$\arg\max_{y \in \mathcal{Y}} P(y|\, \mathbf{x}; \theta) = \arg\max_{y \in \mathcal{Y}} P(\mathbf{x} \,|y; \theta) P(y; \theta)$$

since $P(\mathbf{x}; \theta)$ does not depend on y.

Generative versus Discriminative Models

- In contrast, a discriminative model *directly* specifies the conditional probability of a class y given $\mathbf{x} \Rightarrow P(y|\mathbf{x};\theta)$.
- Therefore, a discriminative model cannot generate the dataset.
- In general, discriminative models give better performance than generative ones when trained on large datasets.

Content

Binary-Valued Features

- Assumption: All features are binary ($x_j \in \{0, 1\}$) and *conditionally independent* given its class y.
- We have

$$\begin{aligned} P(\mathbf{x}, y; \theta) &= P(\mathbf{x} \,|y; \theta) P(y; \theta) \\ &= \prod_{j=1}^{D} P(x_j | y; \theta) P(y; \theta). \end{aligned}$$

- Parameters of the model:

$$\begin{aligned} \theta_k &= P(y = k), \forall k = 1, 2, \ldots, K \\ \theta_{j|k} &= P(x_j = 1 | y = k), \forall j = 1, 2, \ldots, D; \forall k = 1, 2, \ldots, K \end{aligned}$$

- Note that $\theta_K = 1 - \sum_{k=1}^{K-1} \theta_k$, the model has $(K - 1) + DK$ parameters.

Maximum Likelihood Estimation

The likelihood of a dataset of N samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ is

$$L(\theta) = \prod_{i=1}^{N} P(\mathbf{x}_i, y_i) = \prod_{i=1}^{N} \left(\prod_{j=1}^{D} P(x_{ij}|y_i;\theta) P(y_i;\theta) \right)$$
$$= \prod_{i=1}^{N} \left(\prod_{j=1}^{D} \beta_{j|y_i} \theta_{y_i} \right),$$

where

$$\beta_{j|k} = \begin{cases} \theta_{j|k}, & \text{if } x_j = 1 \\ 1 - \theta_{j|k}, & \text{if } x_j = 0. \end{cases}$$

Maximum Likelihood Estimation

Maximum likelihood estimation:

$$\hat{\theta}_k = \frac{\sum_{i=1}^{N} \delta(y_i = k)}{N},$$

$$\hat{\theta}_{j|k} = \frac{\sum_{i=1}^{N} \delta(x_{ij} = 1, y_i = k)}{\sum_{i=1}^{N} \delta(y_i = k)},$$

where $\delta(\cdot)$ is the identity function.

Classification

- Given an object $\mathbf{x}$, its class is determined as

$$y := \hat{k} = \underset{k=1,2,\ldots,K}{\arg\max}\ P(y = k|\,\mathbf{x})$$

$$= \underset{k=1,2,\ldots,K}{\arg\max} \prod_{j=1}^{D} \beta_{j|k}\theta_k$$

- If the loga function is used, we have a linear classifier:

$$y := \hat{k} = \underset{k=1,\ldots,K}{\arg\max} \left(\sum_{j=1}^{D} \log \beta_{j|k} + \log \theta_k \right) .$$

Smoothing

- We need to smooth the model to take into account the case that $\theta_{j|k} = 0$.
- If $\theta_{j|k} = 0, \forall k = 1, 2, \ldots, K$ then

$$P(\mathbf{x}) = \sum_{k=1}^{K} \left(\theta_k \prod_{j=1}^{D} \theta_{j|k} \right) = 0.$$

- So we have

$$P(y = k|\,\mathbf{x}) = \frac{0}{0}, \quad \forall k = 1, 2, \ldots, K.$$

$\Rightarrow$ we cannot construct a classification rule for $\mathbf{x}$.

Smoothing

Laplace smoothing:

$$\hat{\theta}_{j|k} = \frac{\sum_{i=1}^{N} \delta(x_{ij} = 1, y_i = k) + \alpha}{\sum_{i=1}^{N} \delta(y_i = k) + D\alpha},$$

where α is a smoothing term (predefined or estimated by cross-validation).

Example: Sportive Activity

In order to predict whether a person will play an outdoor sportive activity or not, the following weather information is used:

- ngày nắng hay mưa
- nhiệt độ nóng hay mát
- độ ẩm cao hay bình thường
- trời có gió hay không

Sportive Activity

Note that all attributes take binary values. Each object has 4 attributes ($D = 4$):

$$x_1 \in \{\text{nắng}, \text{mưa}\} \equiv \{1, 0\}$$
$$x_2 \in \{\text{nóng}, \text{mát}\} \equiv \{1, 0\}$$
$$x_3 \in \{\text{cao}, \text{bình thường}\} \equiv \{1, 0\}$$
$$x_4 \in \{\text{đúng}, \text{sai}\} \equiv \{1, 0\}$$

Each object is classified into one class $y \in \{\text{có}, \text{không}\}$ ($K = 2$).

Sportive Activity

The training data is as follows ($N = 8$):

Ngoài trời x_1	**Nhiệt độ** x_2	**Độ ẩm** x_3	**Có gió** x_4	**Chơi** y
nắng	nóng	cao	sai	không
nắng	nóng	cao	đúng	không
mưa	nóng	cao	sai	có
mưa	mát	cao	đúng	có
mưa	mát	bình thường	sai	có
mưa	mát	bình thường	đúng	không
nắng	mát	bình thường	sai	có
mưa	nóng	cao	đúng	không

Estimate the parameters of a NB model and predict the chance of sportive activity of a day if it is defined as:

$$\mathbf{x} = (\text{trời nắng}, \text{mát}, \text{độ ẩm cao}, \text{sai})$$

Sportive Activity

$$\hat{\theta}_{1|\text{có}} = \frac{\sum_{i=1}^{8} \delta(x_{i1} = \text{nắng}, y_i = \text{có})}{\sum_{i=1}^{8} \delta(y_i = \text{có})} = \frac{1}{4} = 0.25$$

$$\hat{\theta}_{1|\text{không}} = \frac{\sum_{i=1}^{8} \delta(x_{i1} = \text{nắng}, y_i = \text{không})}{\sum_{i=1}^{8} \delta(y_i = \text{không})} = \frac{2}{4} = 0.5$$

Sportive Activity

$$\hat{\theta}_{1|\text{có}} = \frac{\sum_{i=1}^{8} \delta(x_{i1} = \text{nắng}, y_i = \text{có})}{\sum_{i=1}^{8} \delta(y_i = \text{có})} = \frac{1}{4} = 0.25$$

$$\hat{\theta}_{1|\text{không}} = \frac{\sum_{i=1}^{8} \delta(x_{i1} = \text{nắng}, y_i = \text{không})}{\sum_{i=1}^{8} \delta(y_i = \text{không})} = \frac{2}{4} = 0.5$$

$$\hat{\theta}_{2|\text{có}} = \frac{\sum_{i=1}^{8} \delta(x_{i2} = \text{nóng}, y_i = \text{có})}{\sum_{i=1}^{8} \delta(y_i = \text{có})} = \frac{1}{4} = 0.25$$

$$\hat{\theta}_{2|\text{không}} = \frac{\sum_{i=1}^{8} \delta(x_{i2} = \text{nóng}, y_i = \text{không})}{\sum_{i=1}^{8} \delta(y_i = \text{không})} = \frac{3}{4} = 0.75$$

Sportive Activity

$$\hat{\theta}_{3|\text{có}} = \frac{\sum_{i=1}^{8} \delta(x_{i3} = \text{cao}, y_i = \text{có})}{\sum_{i=1}^{8} \delta(y_i = \text{có})} = \frac{2}{4} = 0.5$$

$$\hat{\theta}_{3|\text{không}} = \frac{\sum_{i=1}^{8} \delta(x_{i3} = \text{cao}, y_i = \text{không})}{\sum_{i=1}^{8} \delta(y_i = \text{không})} = \frac{3}{4} = 0.75$$

Sportive Activity

$$\hat{\theta}_{3|\text{có}} = \frac{\sum_{i=1}^{8} \delta(x_{i3} = \text{cao}, y_i = \text{có})}{\sum_{i=1}^{8} \delta(y_i = \text{có})} = \frac{2}{4} = 0.5$$

$$\hat{\theta}_{3|\text{không}} = \frac{\sum_{i=1}^{8} \delta(x_{i3} = \text{cao}, y_i = \text{không})}{\sum_{i=1}^{8} \delta(y_i = \text{không})} = \frac{3}{4} = 0.75$$

$$\hat{\theta}_{4|\text{có}} = \frac{\sum_{i=1}^{8} \delta(x_{i4} = \text{đúng}, y_i = \text{có})}{\sum_{i=1}^{8} \delta(y_i = \text{có})} = \frac{1}{4} = 0.25$$

$$\hat{\theta}_{4|\text{không}} = \frac{\sum_{i=1}^{8} \delta(x_{i4} = \text{đúng}, y_i = \text{không})}{\sum_{i=1}^{8} \delta(y_i = \text{không})} = \frac{3}{4} = 0.75$$

Sportive Activity

Estimated values of $\theta_{j|k}$ are as follows:

$\theta_{j\|k}$	có	không
nắng	0.25	0.5
nóng	0.25	0.75
cao	0.5	0.75
đúng	0.25	0.75

Sportive Activity

Given a new day $\mathbf{x}$ =(nắng, mát, cao, sai):

$$P(y = \text{có}|\,\mathbf{x}) = \frac{[0.25 \times (1 - 0.25) \times 0.5 \times (1 - 0.25)]\frac{1}{2}}{P(\mathbf{x})}$$
$$= \frac{0.0703125}{2P(\mathbf{x})}$$
$$P(y = \text{không}|\,\mathbf{x}) = \frac{[0.5 \times (1 - 0.75) \times 0.75 \times (1 - 0.75)]\frac{1}{2}}{P(\mathbf{x})}$$
$$= \frac{0.0234375}{2P(\mathbf{x})}.$$

For this $\mathbf{x}$, we predict its $y =$ "có" since

$$P(y = \text{có}|\,\mathbf{x}) > P(y = \text{không}|\,\mathbf{x}).$$

Exercise: Fitting a Naive Bayes Spam Filter by Hand

Consider a Naive Bayes model for spam email classification:

- Vocabulary: 'secret', 'offer', 'low', 'price', 'valued', 'customer', 'today', 'dollar', 'million', 'sports', 'is', 'for', 'play', 'healthy', 'pizza'.
- Example spam messages: "million dollar offer", "secret offer today", "secret is secret"
- Example normal messages: "low price for valued customer", "play secret sports today", "sports is healthy", "low price pizza".

Give the MLEs for

$$\theta_{\text{spam}}, \theta_{\text{secret}|\text{spam}}, \theta_{\text{secret}|\text{non-spam}}, \theta_{\text{sports}|\text{non-spam}}, \theta_{\text{dollar}|\text{spam}}.$$

Content

Recap: Normal Distribution

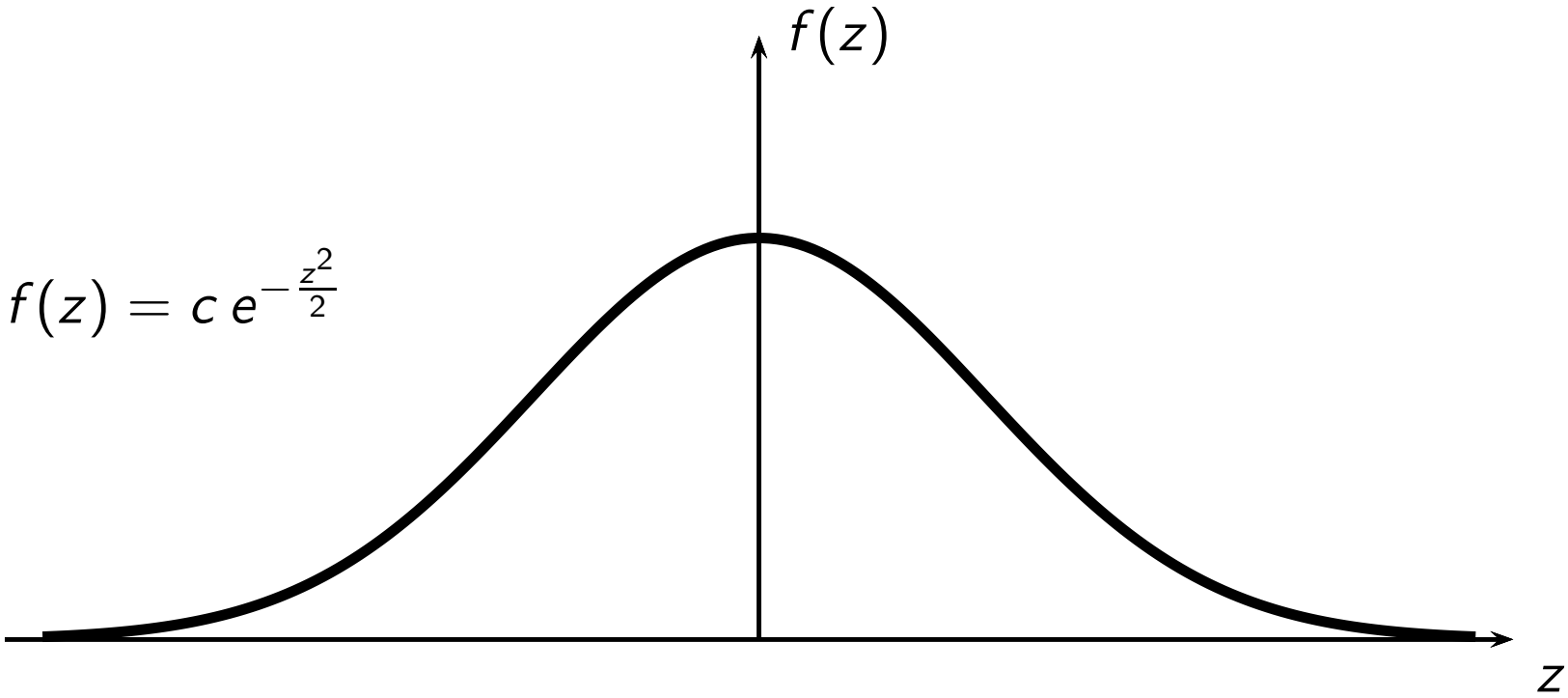
- Normal distribution is one of the most important distributions in probability.
- One of the reason is the central limit theorem: *The sum of a large number of independent random variables has a distribution that is approximately normal.*
- This explains the remarkable fact that the empirical frequencies of so many natural populations exhibit bell-shaped curves.

Standard Normal Distribution

We say that a random variable Z is a standard normal variable with mean 0 and variance 1, denoted as $Z \sim \mathcal{N}(0, 1)$ if the density of Z is given by

$$f(z) = c\, e^{-\frac{z^2}{2}}, z \in \mathbb{R},$$

where c is the normalizing constant.



Standard Normal Distribution

Some remarks on the function $f(z)$:

- It is an even (symmetric) function.
- It decreases very rapidly when z increases.

To find c, we solve the equation:

$$\int_{-\infty}^{\infty} f(z)\, dz = 1.$$

Standard Normal Distribution

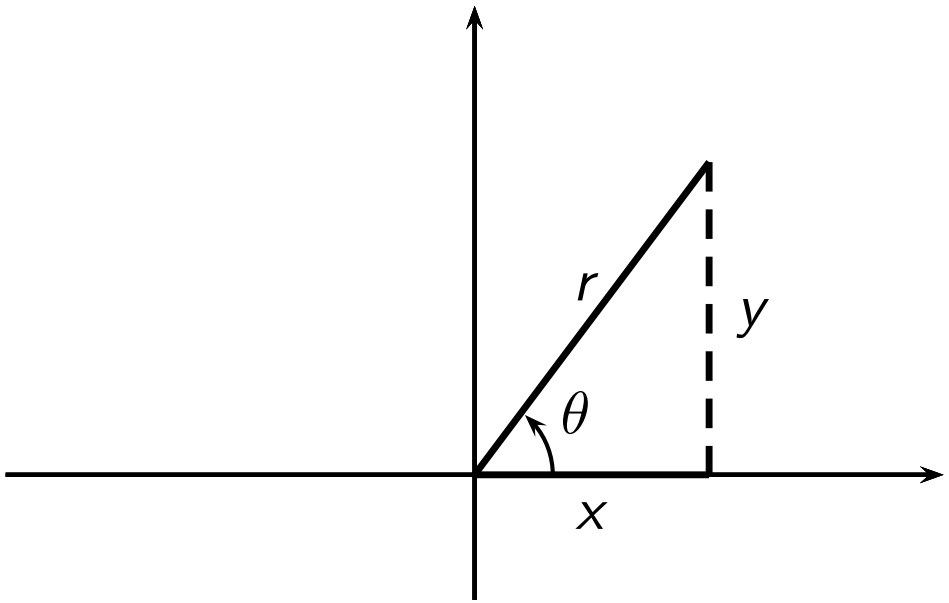
This is not an easy task. We solve the equation as follows:

$$\begin{aligned}
\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx \times \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\,dy &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,dx\,dy \\
&= \int_{0}^{2\pi}\int_{0}^{\infty} e^{-\frac{r^2}{2}}(r)\,dr\,d\theta \\
&= \int_{0}^{2\pi}\left(\int_{0}^{\infty} e^{-\frac{r^2}{2}} r\,dr\right)d\theta = \int_{0}^{2\pi} 1\,d\theta \\
&= 2\pi.
\end{aligned}$$

Standard Normal Distribution

Here, we evaluate the double integral by means of a change of variables from Cartesian coordinates to polar coordinates. We have $x = r\cos\theta$, $y = r\sin\theta$ and

$$J(r,\theta) = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$



The Jacobian of the transformation is $r\cos^2\theta + r\sin^2\theta = r$.

Standard Normal Distribution

Thus,

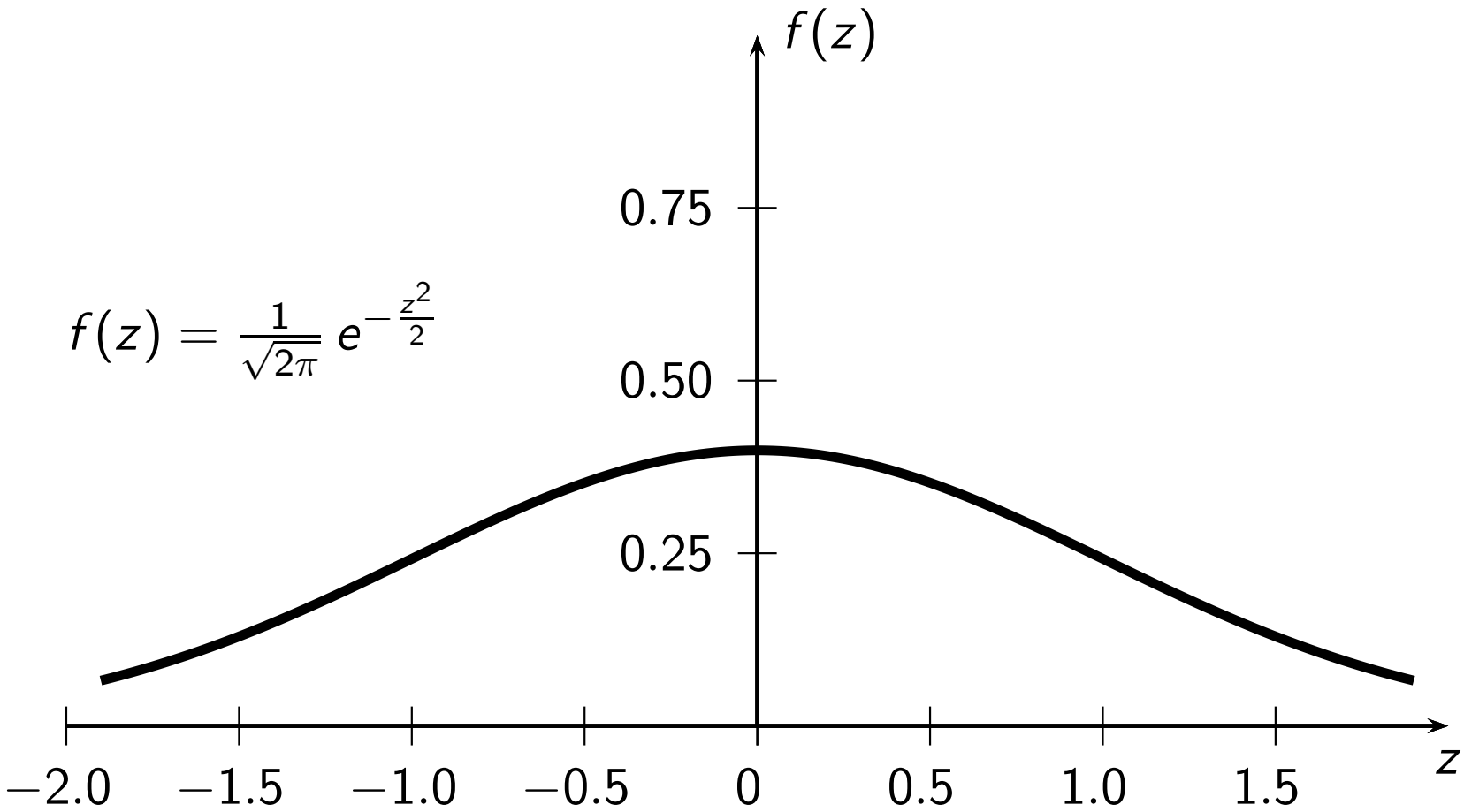
$$\int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\, dz = \sqrt{2\pi},$$

The density of the normally distributed random variable $\mathcal{N}(0, 1)$ is:

$$\boxed{f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, z \in \mathbb{R}.}$$

Standard Normal Distribution

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$



Standard Normal Distribution

If Z is a standard normal variable then

$$\mathbb{E}(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} \, dz = 0.$$

and

$$\text{var}(Z) = \mathbb{E}(Z^2) - [\mathbb{E}(Z)]^2 = \mathbb{E}(Z^2).$$

We have

$$\mathbb{E}(Z^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} \, dz$$

Standard Normal Distribution

Compute the integral by parts:

$$\begin{cases} u &= z \\ dv &= ze^{-\frac{z^2}{2}}\,dz \end{cases} \Rightarrow \begin{cases} du &= dz \\ v &= \int_{-\infty}^{\infty} ze^{-\frac{z^2}{2}}\,dz = -e^{-\frac{z^2}{2}}\Big|_{-\infty}^{\infty} \end{cases}$$

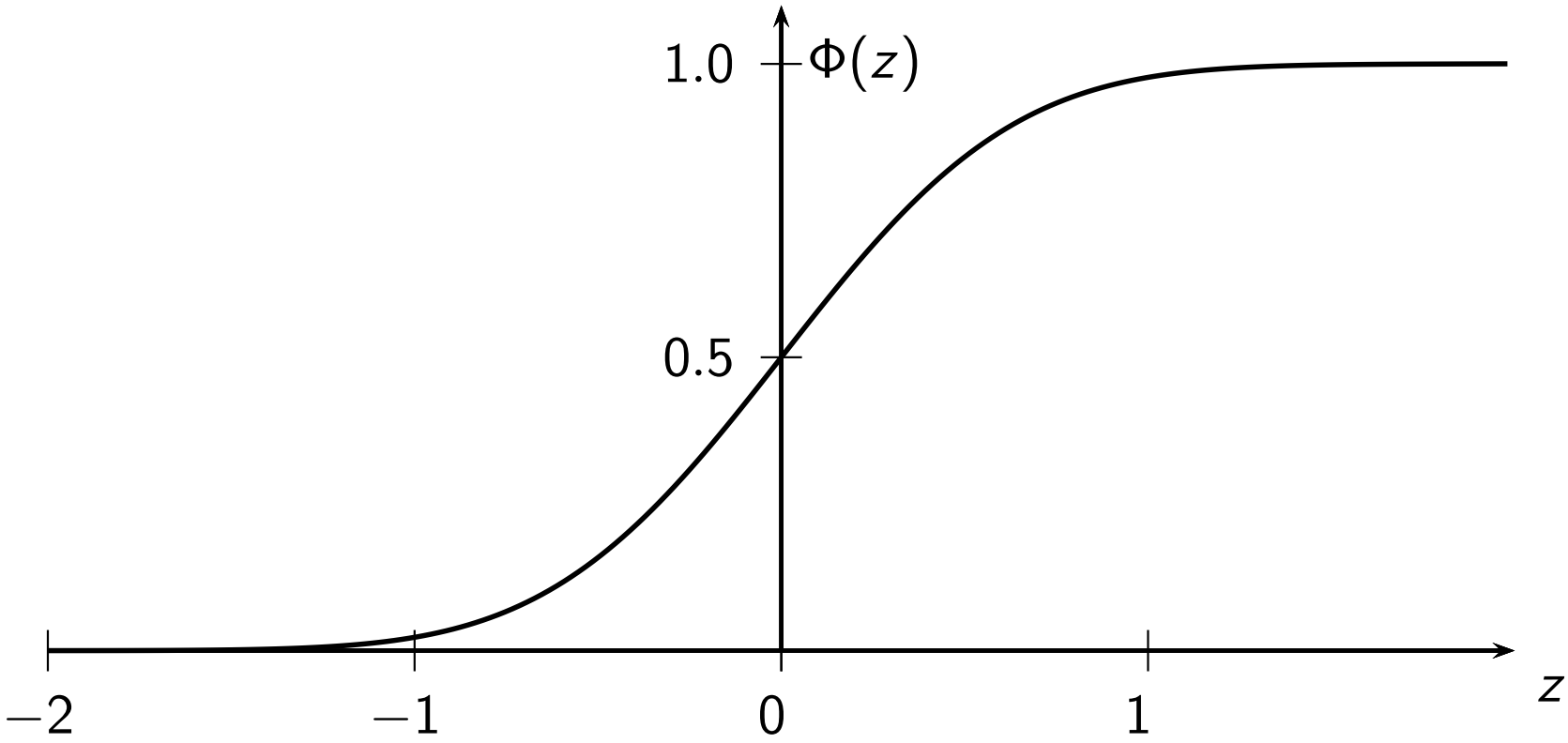
Therefore

$$\begin{aligned} \frac{1}{\sqrt{2\pi}}\int_0^{\infty} z^2 e^{-\frac{z^2}{2}}\,dz &= \frac{1}{\sqrt{2\pi}}\left(-ze^{-\frac{z^2}{2}}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\,dz\right) \\ &= 0 + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\,dz = 1. \end{aligned}$$

Standard Normal Distribution

Let Φ denote the distribution function of the standard normal variable, then

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}}\, dt.$$



Standard Normal Distribution

By symmetry, we have

$$\Phi(-z) = 1 - \Phi(z).$$

If $Z \sim \mathcal{N}(0, 1)$ then for all (a, b):

$$P(a \leq Z \leq b) = \Phi(b) - \Phi(a).$$

Normal Distribution

- A random variable X is normally distributed with parameters (μ, σ^2), denoted as $\mathcal{N}(\mu, \sigma^2)$ if its density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.$$

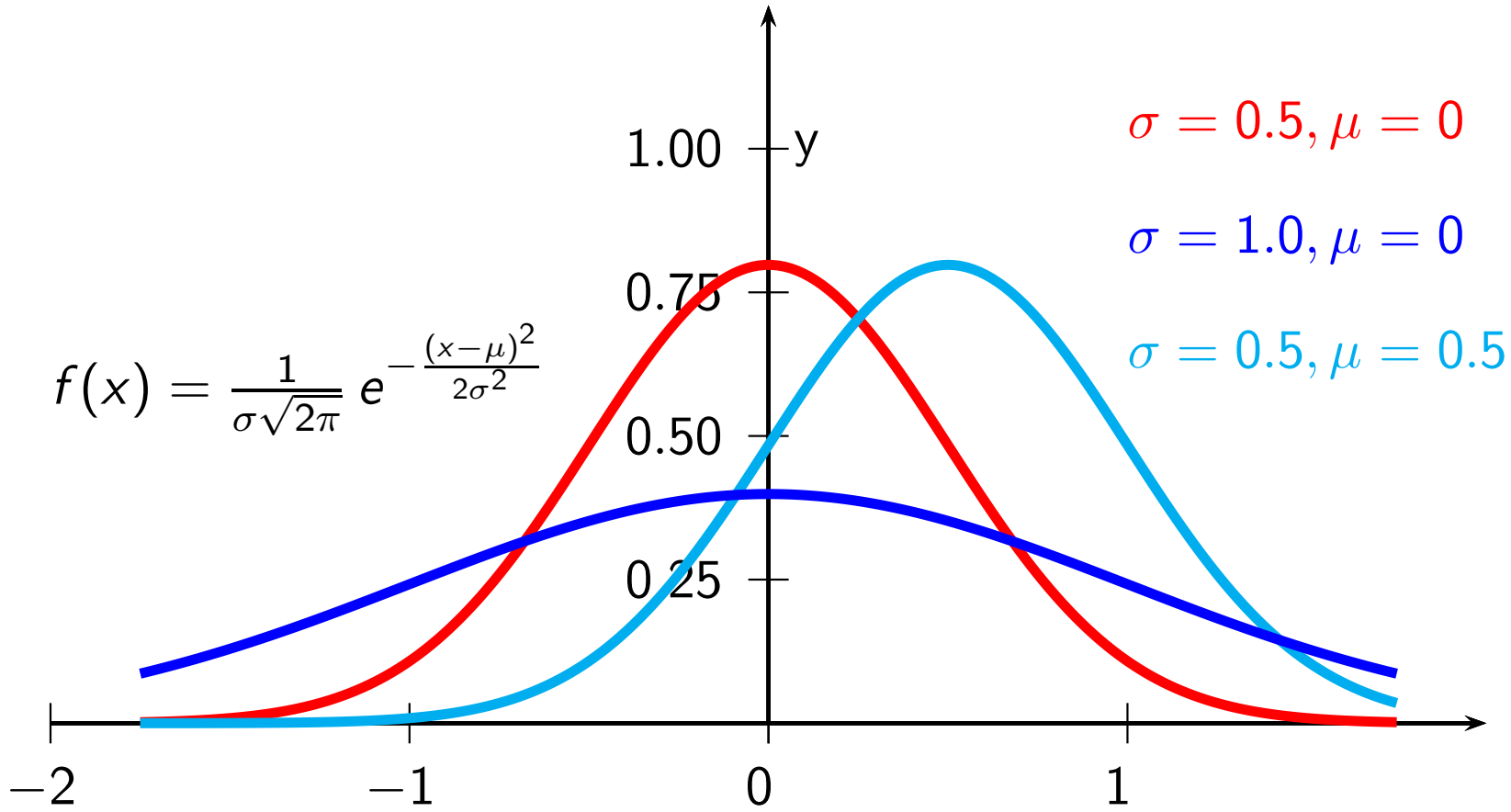
Let $Z = \frac{X-\mu}{\sigma}$, we have

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_{-\infty}^{\infty} f(z)\, dz = 1$$

since Z is the standard normal random variable.

Normal Distribution

Because $X = \mu + \sigma Z$ and due to the linearity of expectation, we have $\mathbb{E}(X) = \mu + \sigma\,\mathbb{E}(Z) = \mu$ and $\text{var}(X) = \sigma^2\,\text{var}(Z) = \sigma^2$.



Normal Distribution

Some properties

- If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$ then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.
- If $X \sim \mathcal{N}(\mu, \sigma^2)$ then the distribution function of X can be computed via the standard normal distribution:

$$F_X(a) = P(X \leq a) = P\left(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right)$$
$$= P\left(Z \leq \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right).$$

- If $X \sim \mathcal{N}(\mu, \sigma^2)$ then for all intervals (a, b), we have

$$P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right)$$
$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Gaussian Naive Bayes Model

- If $x_j \in \mathbb{R}$ we can suppose that $x_j|y = k$ is distributed normally with mean μ_{jk} and variance σ_{jk}^2. We have

$$P(x_j|y = k) = \frac{1}{\sqrt{2\pi\sigma_{jk}^2}} \exp\left(-\frac{(x_j - \mu_{jk})^2}{2\sigma_{jk}^2}\right).$$

Gaussian Naive Bayes Model

- Parameters of the model:

$$\theta_k = P(y = k), \forall k = 1, 2, \ldots, K$$
$$\theta_{j|k} = P(x_j | y = k), \forall k = 1, 2, \ldots, K; \forall j = 1, 2, \ldots, D.$$

- In other words, for each class k, we have priors θ_k and D pairs of parameters for D normal distributions $\theta_{j|k} := (\mu_{jk}, \sigma^2_{jk})$.
- The number of parameters of the model is $(K - 1) + 2DK$.

Maximum Likelihood Estimation

We have a training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$. The conditional likelihood of each class k on this set is

$$L(\theta|k) = \prod_{i=1}^{N} P(\mathbf{x}_i \,|y_i = k) = \prod_{i=1}^{N} \left(\prod_{j=1}^{D} P(x_{ij}|y_i) \right).$$

The log-likelihood is

$$\ell(\theta|k) = \sum_{i=1}^{N} \sum_{j=1}^{D} \log P(x_{ij}|y_i).$$

Maximum Likelihood Estimation

Since

$$\log P(x_{ij}|y_i = k) = \log \frac{1}{\sqrt{2\pi\sigma_{jk}^2}} - \frac{(x_{ij} - \mu_{jk})^2}{2\sigma_{jk}^2},$$

we have

$$\ell(\theta|k) = \sum_{i=1}^{N}\sum_{j=1}^{D}\left(\log \frac{1}{\sqrt{2\pi\sigma_{jk}^2}} - \frac{(x_{ij} - \mu_{jk})^2}{2\sigma_{jk}^2}\right).$$

Maximum Likelihood Estimation

Therefore

$$\begin{aligned}\frac{\partial \ell}{\partial \mu_{jk}} &= \sum_{i=1}^{N_k} \frac{1}{\sigma_{jk}^2}(x_{ij} - \mu_{jk}) \\ &= \frac{1}{\sigma_{jk}^2}\left(\sum_{i=1}^{N_k} x_{ij} - N_k \mu_{jk}\right).\end{aligned}$$

Thus, the maximum likelihood estimate of μ_{jk} is

$$\boxed{\hat{\mu}_{jk} = \frac{1}{N_k}\sum_{i=1}^{N_k} x_{ij}, \text{where } y_i = k}$$

Maximum Likelihood Estimation

Let $t_{jk} = \sigma_{jk}^2$, we have

$$\begin{aligned}
\frac{\partial \ell}{\partial t_{jk}} &= \sum_{i=1}^{N_k} \frac{\partial \ell}{\partial t_{jk}} \left(\log \frac{1}{\sqrt{2\pi t_{jk}}} - \frac{(x_{ij} - \mu_{jk})^2}{2t_{jk}} \right) \\
&= \sum_{i=1}^{N_k} \left(-\frac{1}{2t_{jk}} + \frac{(x_{ij} - \mu_{jk})^2}{2t_{jk}^2} \right) \\
&= \frac{1}{2t_{jk}^2} \left(-N_k t_{jk} + \sum_{i=1}^{N_k} (x_{ij} - \mu_{jk})^2 \right).
\end{aligned}$$

So, the maximum likelihood estimate of σ_{jk}^2 is

$$\boxed{\hat{\sigma}_{jk}^2 = \frac{1}{N_k} \sum_{i=1}^{N_k} (x_{ij} - \hat{\mu}_{jk})^2.}$$

Classification

Once we have estimators of parameters, we may predict the class $\hat{y}$ of a data point $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ as follows:

$$\hat{y} = \underset{k=1,2,\ldots,K}{\arg\max} \log P(y = k | \mathbf{x}),$$

where

$$\begin{aligned}
\log P(y = k | \mathbf{x}) &\propto \log \left[P(y = k) P(\mathbf{x} | y = k) \right] \\
&\propto \log \left[P(y = k) \prod_{j=1}^{D} P(x_j | y = k) \right] \\
&\propto \log P(y = k) + \sum_{j=1}^{D} \log P(x_j | y = k).
\end{aligned}$$

Classification

That is,

$$\hat{y} = \underset{k=1,2,\ldots,K}{\arg\max} \left[\log \hat{\theta}_k + \sum_{j=1}^{D} \left(\log \frac{1}{\sqrt{2\pi\hat{\sigma}_{jk}^2}} - \frac{(x_j - \hat{\mu}_{jk})^2}{2\hat{\sigma}_{jk}^2} \right) \right]$$

Example: Breast Cancer

- Áp dụng mô hình GNB trong chẩn đoán y học trên một bộ dữ liệu về bệnh nhân ung thư vú của Đại học Wisconsin–Madison, Hoa Kỳ.[1].
- Để dự báo một bệnh nhân có mắc phải bệnh ung thư vú hay không, người ta lấy mẫu sinh thiết (FNA–Fine Needle Aspiration) của khối u và phân tích mẫu này.
- Lấy một lát cắt của mẫu để soi dưới kính hiển vi, quét và ghi lại mẫu dưới dạng các khung ảnh số của các nhân tế bào.
- Sau khi đã cô lập các nhân tế bào, ta tiến hành tính toán 10 đặc tính của nhân, đo kích thước, hình dạng, kết cấu của chúng.
- Với mỗi đặc trưng này, ta tính toán *giá trị trung bình*, *độ lệch chuẩn*, *các giá trị cực trị*.

[1] Wisconsin Diagnostic Breast Cancer (WDBC), `http://www.cs.wisc.edu/~olvi/uwmp/cancer.html`

Example: Breast Cancer

- Có 30 đặc trưng giá trị thực cho mỗi mẫu.
- Tập huấn luyện có 569 mẫu dữ liệu, trong đó có 357 mẫu u lành tính, 212 mẫu u ác tính.
- Với mỗi nhân tế bào, người ta tính 10 đặc trưng sau:

STT.	Đặc trưng	Giải thích
0.	radius	trung bình các khoảng cách từ trung tâm tới các điểm trên chu vi
1.	texture	kết cấu, là độ lệch chuẩn của các giá trị thang xám
2.	perimeter	chu vi
3.	area	diện tích
4.	smoothness	độ trơn, là độ biến đổi cục bộ theo các độ dài bán kính
5.	compactness	độ đặc, tính bởi perimeter^2 / area - 1.0
6.	concavity	độ lõm
7.	concave points	số phần lõm của các đường viền
8.	symmetry	độ đối xứng
9.	fractal dimension	số chiều fractal

Example: Breast Cancer

- Mỗi mẫu cần chẩn đoán có 30 đặc trưng:
  - Đặc trưng thứ nhất là trung bình của radius
  - Đặc trưng thứ 11 là độ lệch chuẩn của radius
  - Đặc trưng thứ 21 là radius lớn nhất, được tính bởi giá trị trung bình của 3 giá trị lớn nhất.
- Hai mẫu ví dụ được gán các lớp tương ứng là *ác tính* (*M*–malignant) và *lành tính* (*B*–benign).
  1. M, 17.99, 10.38, 122.8, 1001, 0.1184, 0.2776, 0.3001, 0.1471, 0.2419, 0.07871, 1.095, 0.9053, 8.589, 153.4, 0.006399, 0.04904, 0.05373, 0.01587, 0.03003, 0.006193, 25.38, 17.33, 184.6, 2019, 0.1622, 0.6656, 0.7119, 0.2654, 0.4601, 0.1189
  2. B, 7.76, 24.54, 47.92, 181, 0.05263, 0.04362, 0, 0, 0.1587, 0.05884, 0.3857, 1.428, 2.548, 19.15, 0.007189, 0.00466, 0, 0, 0.02676, 0.002783, 9.456, 30.37, 59.16, 268.6, 0.08996, 0.06444, 0, 0, 0.2871, 0.07039

Example: Breast Cancer

- Các xác suất tiên nghiệm cho các lớp M và B tương ứng là

$$\hat{\theta}_M = \frac{357}{569} \approx 0.6274$$
$$\hat{\theta}_B = \frac{212}{569} \approx 0.3726.$$

- Khi xây dựng mô hình, ta có hai lựa chọn là chỉ sử dụng 10 đặc trưng đầu tiên hoặc sử dụng toàn bộ 30 đặc trưng.
- Với mỗi lựa chọn đó, ta xây dựng hai mô hình:
  - Dùng chung $\sigma_j = \sigma_{j|M} = \sigma_{j|B}$ cho các lớp M và B
  - Dùng riêng $\sigma_{j|M} \neq \sigma_{j|B}$.

Example: Breast Cancer

Nếu sử dụng 10 đặc trưng đầu tiên và dùng chung σ_j:

j	μ_M	μ_B	σ
0.	17.46283	12.146524	3.524049
1.	21.604906	17.914762	4.301036
2.	115.365377	78.075406	24.298981
3.	978.376415	462.790196	351.914129
4.	0.102898	0.092478	0.014064
5.	0.145188	0.080085	0.052813
6.	0.160775	0.046058	0.07972
7.	0.08799	0.025717	0.038803
8.	0.192909	0.174186	0.027414
9.	0.06268	0.062867	0.00706

Độ chính xác phân loại trên tập huấn luyện 92.79%.

Example: Breast Cancer

Nếu sử dụng 10 đặc trưng đầu tiên và dùng σ_j cho từng lớp:

j	μ_M	μ_B	σ_M	σ_B
0.	17.46283	12.146524	3.203971	1.780512
1.	21.604906	17.914762	3.77947	3.995125
2.	115.365377	78.075406	21.854653	11.807438
3.	978.376415	462.790196	367.937978	134.287118
4.	0.102898	0.092478	0.012608	0.013446
5.	0.145188	0.080085	0.053987	0.03375
6.	0.160775	0.046058	0.075019	0.043442
7.	0.08799	0.025717	0.034374	0.015909
8.	0.192909	0.174186	0.027638	0.024807
9.	0.06268	0.062867	0.007573	0.006747

Độ chính xác phân loại trên tập huấn luyện 91.38%.

Example: Breast Cancer

- Sử dụng toàn bộ 30 đặc trưng và σ_j không phụ thuộc vào lớp thì thì ta thấy mô hình cho kết quả chính xác 92.97% trên tập dữ liệu huấn luyện.
- Nếu sử dụng toàn bộ 30 đặc trưng và σ_j phụ thuộc vào lớp thì độ chính xác trên tập huấn luyện là 94.02%.

Example: Iris Flowers

- Tập dữ liệu về hoa Iris[2] nổi tiếng trong lĩnh vực nhận dạng.
- Xuất hiện trong bài báo của Ronald Fisher năm 1936, ngày nay vẫn được dùng thường xuyên.
- Tập huấn luyện: 130 mẫu, tập kiểm tra: 20 mẫu

Đặc trưng	Lớp
độ dài của lá đài	`Setosa`
độ rộng của lá đài	`Versicolour`
độ dài của cánh hoa	`Virginica`
độ rộng của cánh hoa	

[2] `http://archive.ics.uci.edu/ml/datasets/Iris`

Example: Iris Flowers

Ước lượng các xác suất tiên nghiệm:

Lớp	$\log(\hat{\theta})$
`setosa`	-1.106
`versicolor`	-1.083
`virginica`	-1.106

Ước lượng trung bình và phương sai của mỗi đặc trưng:

Lớp	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{\mu}_3$	$\hat{\sigma}_3$	$\hat{\mu}_4$	$\hat{\sigma}_4$
`setosa`	5.013	0.354	3.402	0.390	1.462	0.152	0.241	0.109
`versicolor`	5.902	0.520	2.759	0.311	4.229	0.483	1.313	0.204
`virginica`	6.560	0.638	2.969	0.334	5.532	0.566	2.016	0.282

Độ chính xác của mô hình:

- Trên tập kiểm tra: 100% (20/20)
- Trên tập huấn luyện: 95.38% (124/130)

Example: Iris Flowers

- Nếu mô hình được huấn luyện với σ_j không phụ thuộc vào lớp, thì ta có các ước lượng của σ_j ứng với các đặc trưng như sau:

$$\hat{\sigma}_1 = 0.815$$
$$\hat{\sigma}_2 = 0.436$$
$$\hat{\sigma}_3 = 1.752$$
$$\hat{\sigma}_4 = 0.759$$

- Độ chính xác của mô hình này giảm đi so với mô hình trước:
  - Trên tập kiểm tra: 85.00% (17/20)
  - Trên tập huấn luyện: 87.69% (114/130)

Summary

- Some important concepts of supervised classification
- Binary-valued and real-valued features in naive Bayes models
- Naive Bayes assumption: conditionally independent features given a class

$$(x_j|y = k) \text{ is indenpendent to } (x_v|y = k), \forall j \neq v$$

- Training: Maximum Likelihood Estimation – solve an optimization problem
- Examples: Sportive Activity, Spam Email, Breast Cancer, Iris Flowers