Naïve Bayes classification

Probability theory

Random variable: a variable whose possible values are numerical outcomes of a random phenomenon.

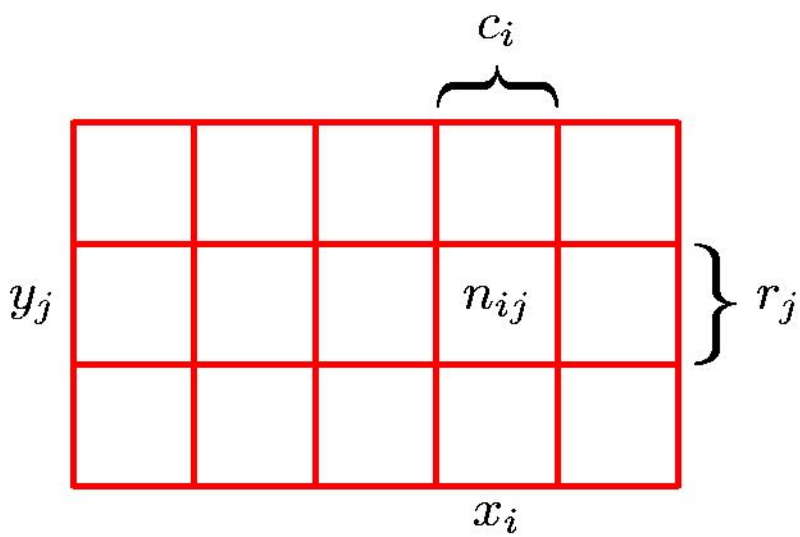
Examples: A person's height, the outcome of a coin toss

Distinguish between discrete and continuous variables.

The distribution of a discrete random variable:
The probabilities of each value it can take.
Notation: $P(X = x_i)$.
These numbers satisfy:

$$\sum_i P(X = x_i) = 1$$

Probability theory



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

Joint Probability

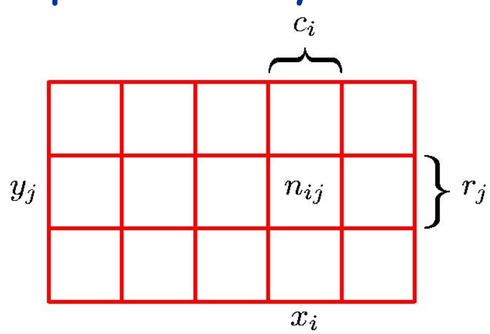
$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Probability theory

A joint probability distribution for two variables is a table.

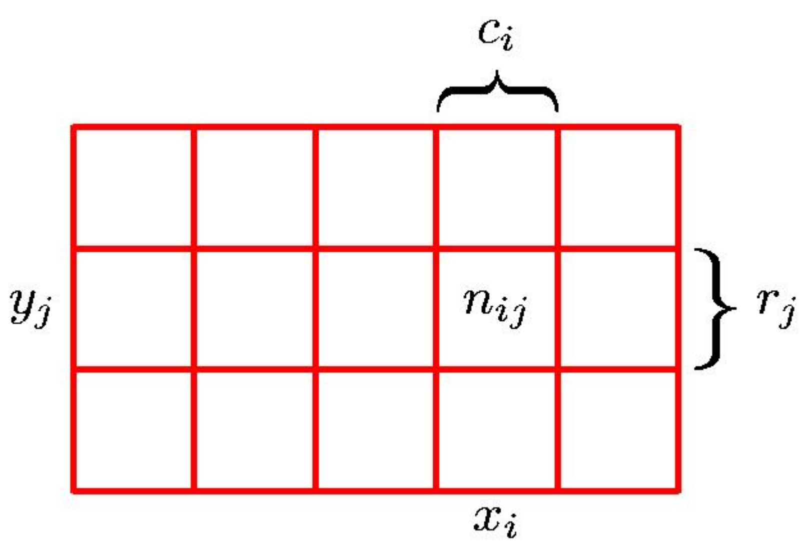


If the two variables are binary, how many parameters does it have?

Let's consider now the joint probability of d variables $P(X_1,\dots,X_d)$.

How many parameters does it have if each variable is binary?

Probability theory



Marginalization:

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned}$$

The Rules of Probability

Marginalization $$p(X) = \sum_Y p(X,Y)$$

Product Rule $$p(X,Y) = p(Y|X)p(X)$$

Independence: X and Y are independent if $P(Y|X) = P(Y)$

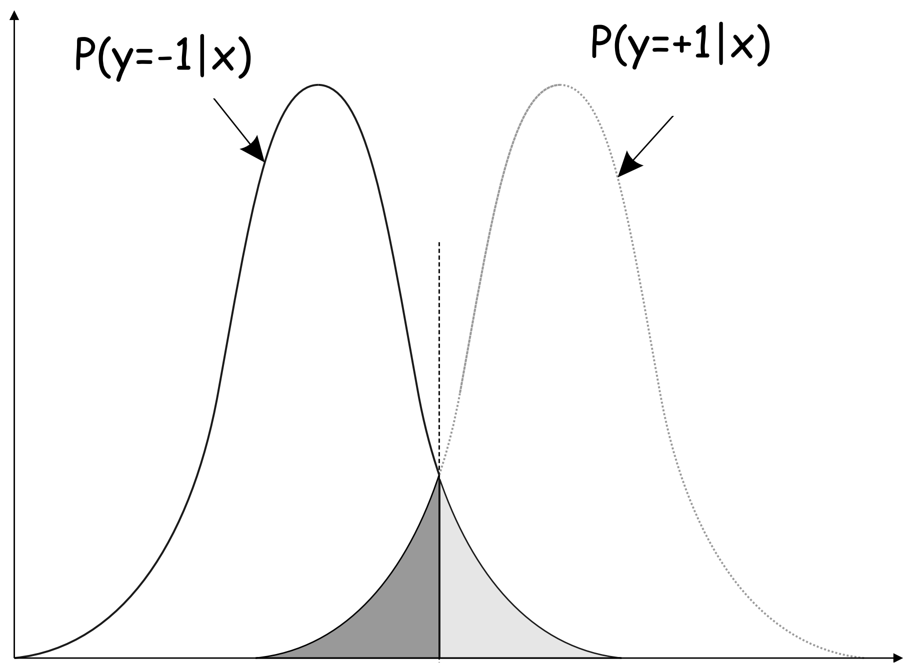
This implies $P(X,Y) = P(X)\,P(Y)$

Using probability in learning

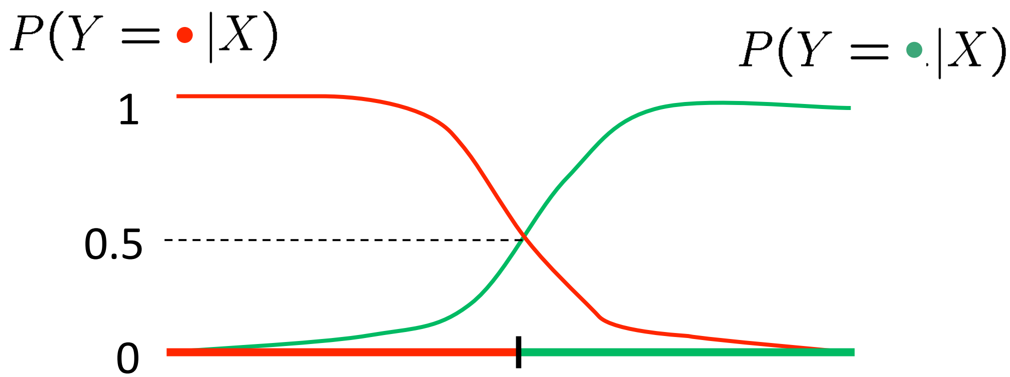
Suppose we have access to $P(y \mid \mathbf{x})$, we would classify according to $\mathrm{argmax}_y\, P(y \mid \mathbf{x})$.

This is called the Bayes-optimal classifier.

What is the error of such a classifier?



Using probability in learning



Some classifiers model P(Y | X) directly: discriminative learning

However, it's usually easier to model P(X | Y) from which we can get P(Y | X) using Bayes rule:

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

This is called generative learning

Maximum likelihood

Fit a probabilistic model P(x | θ) to data

- Estimate θ

Given independent identically distributed (i.i.d.) data $\mathbf{X} = (x_1, x_2, \ldots, x_n)$

- Likelihood

$$P(\mathbf{X}|\theta) = P(x_1|\theta)P(x_2|\theta), \ldots, P(x_n|\theta)$$

- Log likelihood

$$\ln P(\mathbf{X}|\theta) = \sum_{i=1}^{n} \ln P(x_i|\theta)$$

Maximum likelihood solution: parameters θ that maximize ln P(X | θ)

Example

Example: coin toss

Estimate the probability p that a coin lands "Heads" using the result of n coin tosses, h of which resulted in heads.

The likelihood of the data: $P(\mathbf{X}|\theta) = p^h(1-p)^{n-h}$

Log likelihood: $\ln P(\mathbf{X}|\theta) = h \ln p + (n-h)\ln(1-p)$

Taking a derivative and setting to 0:

$$\frac{\partial \ln P(\mathbf{X}|\theta)}{\partial p} = \frac{h}{p} - \frac{(n-h)}{(1-p)} = 0$$

$$\Rightarrow \quad p = \frac{h}{n}$$

Bayes' rule

From the product rule:

$$P(Y, X) = P(Y \mid X)\, P(X)$$

and:

$$P(Y, X) = P(X \mid Y)\, P(Y)$$

Therefore:

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

This is known as Bayes' rule

Bayes' rule

likelihood prior

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

posterior

posterior ∝ likelihood × prior

P(X) can be computed as:

$$P(X) = \sum_{Y} P(X|Y)P(Y)$$

But is not important for inferring a label

Maximum a-posteriori and maximum likelihood

The maximum a posteriori (MAP) rule:

$$y_{MAP} = \arg\max_Y P(Y|X) = \arg\max_Y \frac{P(X|Y)P(Y)}{P(X)} = \arg\max_Y P(X|Y)P(Y)$$

If we ignore the prior distribution or assume it is uniform we obtain the maximum likelihood rule:

$$y_{ML} = \arg\max_Y P(X|Y)$$

A classifier that has access to P(Y|X) is a Bayes optimal classifier.

Classification with Bayes' rule

We would like to model P(X | Y), where X is a feature vector, and Y is its associated label.

Task: Predict whether or not a picnic spot is enjoyable

Training Data: $X = (X_1 \quad X_2 \quad X_3 \quad \ldots \quad \ldots \quad X_d) \quad Y$

n rows

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

How many parameters?

Prior: P(Y) k-1 if k classes

Likelihood: P(X | Y) $(2^d - 1)k$ for binary features

Naïve Bayes classifier

We would like to model P(X | Y), where X is a feature vector, and Y is its associated label.

Simplifying assumption: conditional independence: given the class label the features are independent, i.e.

$$P(\mathbf{X}|Y) = P(x_1|Y)P(x_2|Y), \ldots, P(x_d|Y)$$

How many parameters now?

Naïve Bayes classifier

We would like to model P(X | Y), where X is a feature vector, and Y is its associated label.

Simplifying assumption: conditional independence: given the class label the features are independent, i.e.

$$P(\mathbf{X}|Y) = P(x_1|Y)P(x_2|Y), \ldots, P(x_d|Y)$$

How many parameters now? dk + k - 1

Naïve Bayes classifier

Naïve Bayes decision rule:

$$y_{NB} = \arg\max_Y P(\mathbf{X}|Y)P(Y) = \arg\max_Y \prod_{i=1}^{d} P(x_i|Y)P(Y)$$

If conditional independence holds, NB is an optimal classifier!

Training a Naïve Bayes classifier

Training data: Feature matrix X (n x d) and labels $y_1,\ldots y_n$

Maximum likelihood estimates:

Class prior: $$\hat{P}(y) = \frac{|\{i : y_i = y\}|}{n}$$

Likelihood: $$\hat{P}(x_i|y) = \frac{\hat{P}(x_i, y)}{\hat{P}(y)} = \frac{|\{i : X_{ij} = x_i, y_i = y\}|/n}{|\{i : y_i = y\}|/n}$$

Example

Email classification

Suppose our vocabulary contains three words a, b and c, and we use a multivariate Bernoulli model for our e-mails, with parameters

$$\boldsymbol{\theta}^{\oplus} = (0.5, 0.67, 0.33) \qquad \boldsymbol{\theta}^{\ominus} = (0.67, 0.33, 0.33)$$

This means, for example, that the presence of b is twice as likely in spam (+), compared with ham.

The e-mail to be classified contains words a and b but not c, and hence is described by the bit vector $\mathbf{x} = (1, 1, 0)$. We obtain likelihoods

$$P(\mathbf{x}|\oplus) = 0.5 \cdot 0.67 \cdot (1 - 0.33) = 0.222 \qquad P(\mathbf{x}|\ominus) = 0.67 \cdot 0.33 \cdot (1 - 0.33) = 0.148$$

The ML classification of $\mathbf{x}$ is thus spam.

Example

Email classification: training data

E-mail	a?	b?	c?	Class
e_1	0	1	0	$+$
e_2	0	1	1	$+$
e_3	1	0	0	$+$
e_4	1	1	0	$+$
e_5	1	1	0	$-$
e_6	1	0	1	$-$
e_7	1	0	0	$-$
e_8	0	0	0	$-$

What are the parameters of the model?

Example

Email classification: training data

E-mail	$a?$	$b?$	$c?$	Class
e_1	0	1	0	+
e_2	0	1	1	+
e_3	1	0	0	+
e_4	1	1	0	+
e_5	1	1	0	−
e_6	1	0	1	−
e_7	1	0	0	−
e_8	0	0	0	−

What are the parameters of the model?

$$\hat{P}(y) = \frac{|\{i : y_i = y\}|}{n}$$

$$\hat{P}(x_i|y) = \frac{\hat{P}(x_i, y)}{\hat{P}(y)} = \frac{|\{i : X_{ij} = x_i, y_i = y\}|/n}{|\{i : y_i = y\}|/n}$$

Example

Email classification: training data

E-mail	a?	b?	c?	Class
e_1	0	1	0	$+$
e_2	0	1	1	$+$
e_3	1	0	0	$+$
e_4	1	1	0	$+$
e_5	1	1	0	$-$
e_6	1	0	1	$-$
e_7	1	0	0	$-$
e_8	0	0	0	$-$

What are the parameters of the model?

P(+) = 0.5, P(-) = 0.5

$$\hat{P}(y) = \frac{|\{i : y_i = y\}|}{n}$$

P(a|+) = 0.5,   P(a|-) = 0.75
P(b|+) = 0.75, P(b|-)= 0.25
P(c|+) = 0.25, P(c|-)= 0.25

$$\hat{P}(x_i|y) = \frac{\hat{P}(x_i, y)}{\hat{P}(y)} = \frac{|\{i : X_{ij} = x_i, y_i = y\}|/n}{|\{i : y_i = y\}|/n}$$

Comments on Naïve Bayes

Usually features are not conditionally independent, i.e.

$$P(\mathbf{X}|Y) \neq P(x_1|Y)P(x_2|Y), \ldots, P(x_d|Y)$$

And yet, one of the most widely used classifiers. Easy to train!

It often performs well even when the assumption is violated.

Domingos, P., & Pazzani, M. (1997). Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. *Machine Learning*. 29, 103-130.

When there are few training examples

What if you never see a training example where x_1=a when y=spam?

$P(x \mid spam) = P(a \mid spam)\, P(b \mid spam)\, P(c \mid spam) = 0$

What to do?

When there are few training examples

What if you never see a training example where x_1=a when y=spam?

P(x | spam) = P(a | spam) P(b | spam) P(c | spam) = 0

What to do?

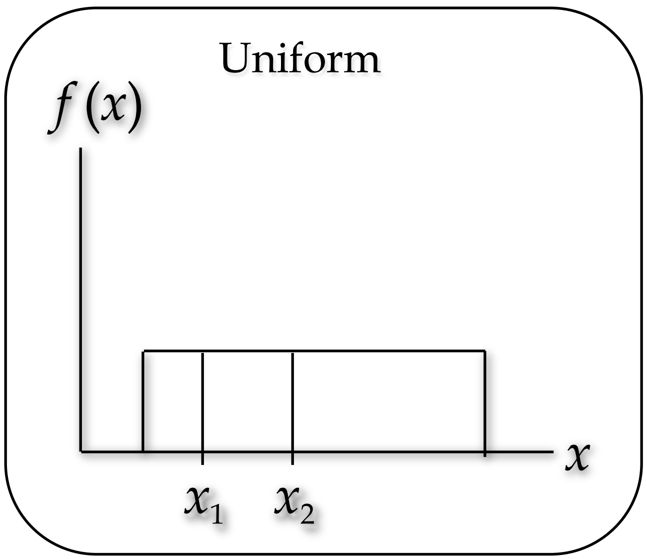
Add “virtual” examples for which x_1=a when y=spam.

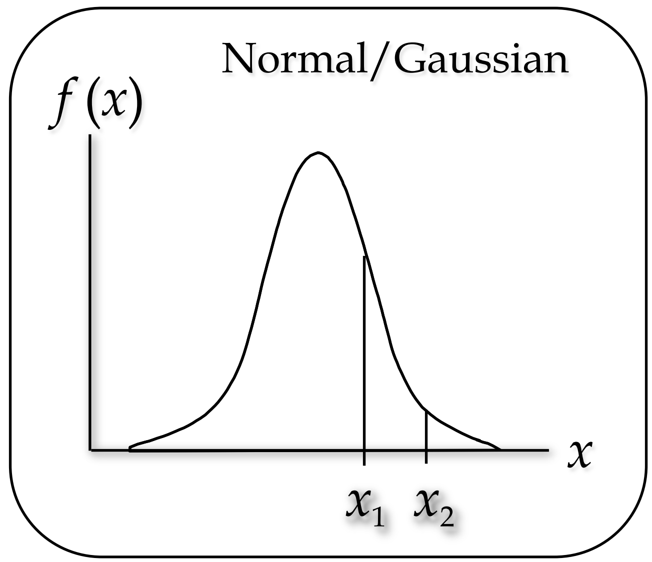
Naïve Bayes for continuous variables

Need to talk about continuous distributions!

Continuous Probability Distributions

The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the <u>area under the graph</u> of the <u>probability density function</u> between x_1 and x_2.





Expectations

Discrete variables

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

Continuous variables

$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x$$

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

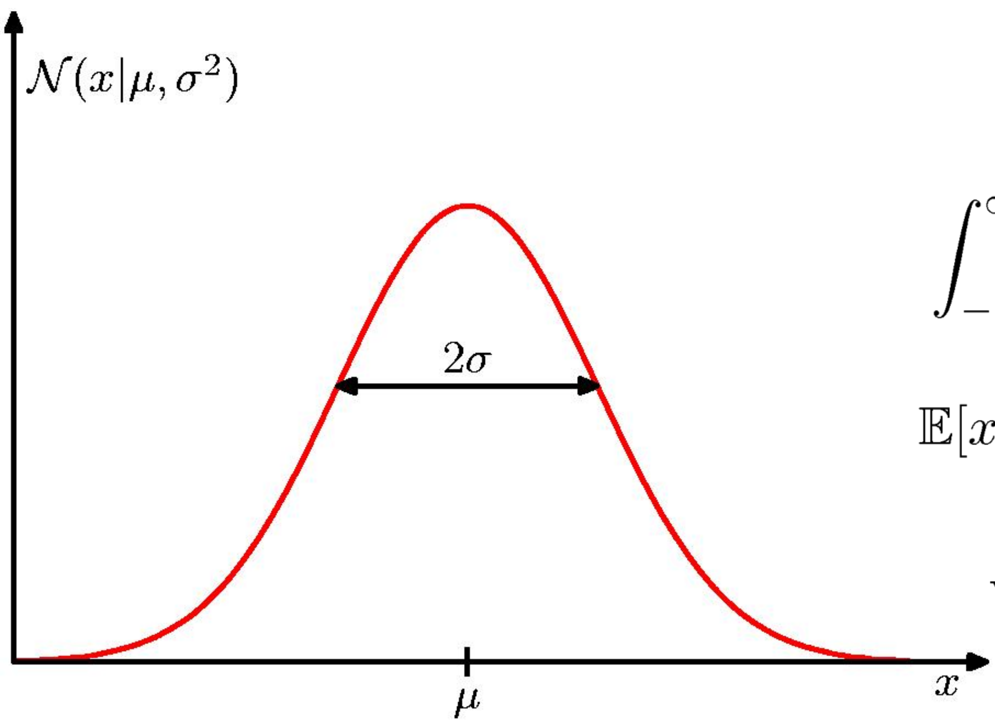
Conditional expectation (discrete)

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

Approximate expectation (discrete and continuous)

The Gaussian (normal) distribution

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



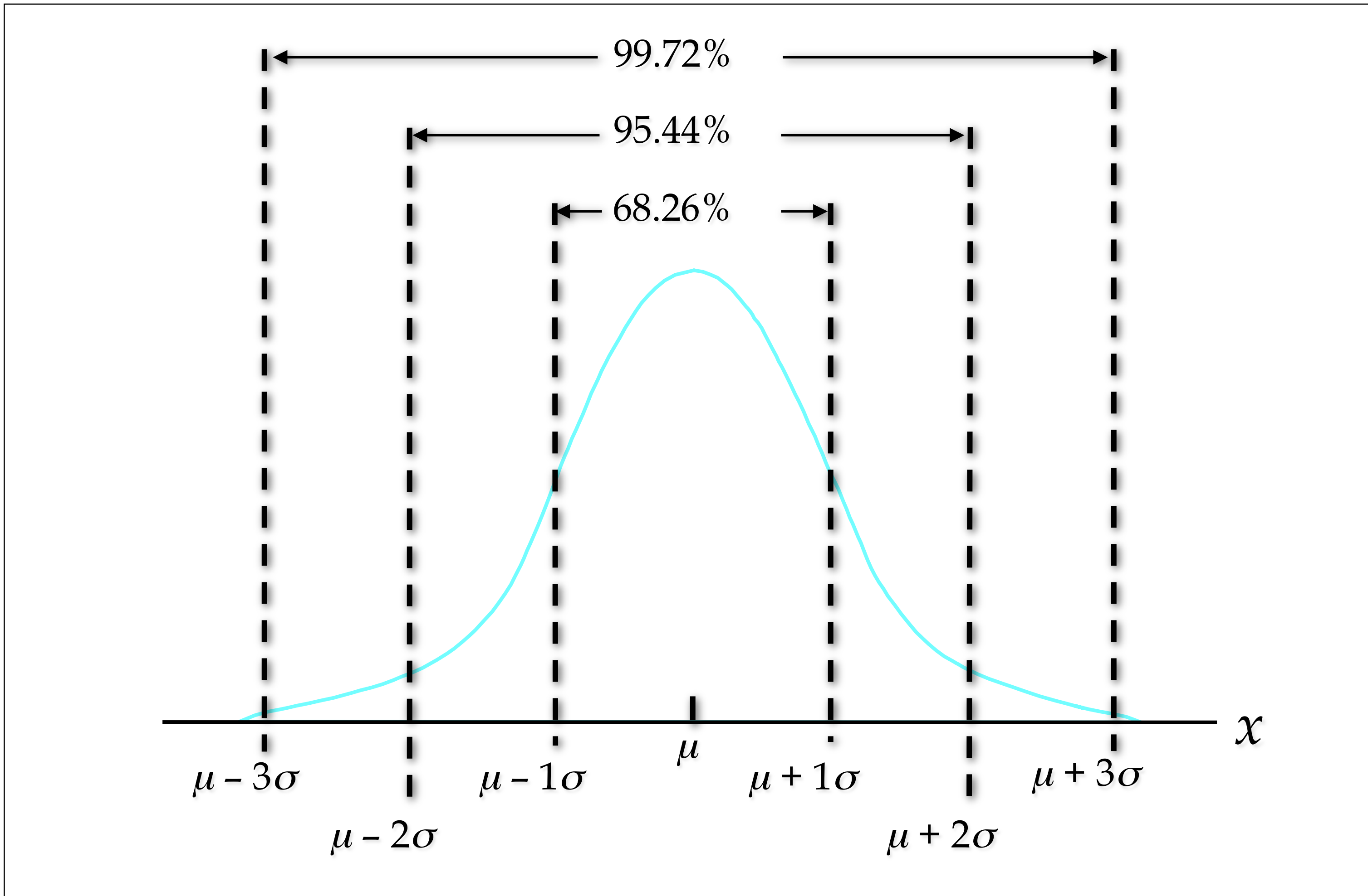
$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\, \mathrm{d}x = 1$$

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x \,\mathrm{d}x = \mu$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

Properties of the Gaussian distribution



Standard Normal Distribution

- A random variable having a normal distribution with a mean of 0 and a standard deviation of 1 is said to have a standard normal probability distribution.

Standard Normal Probability Distribution

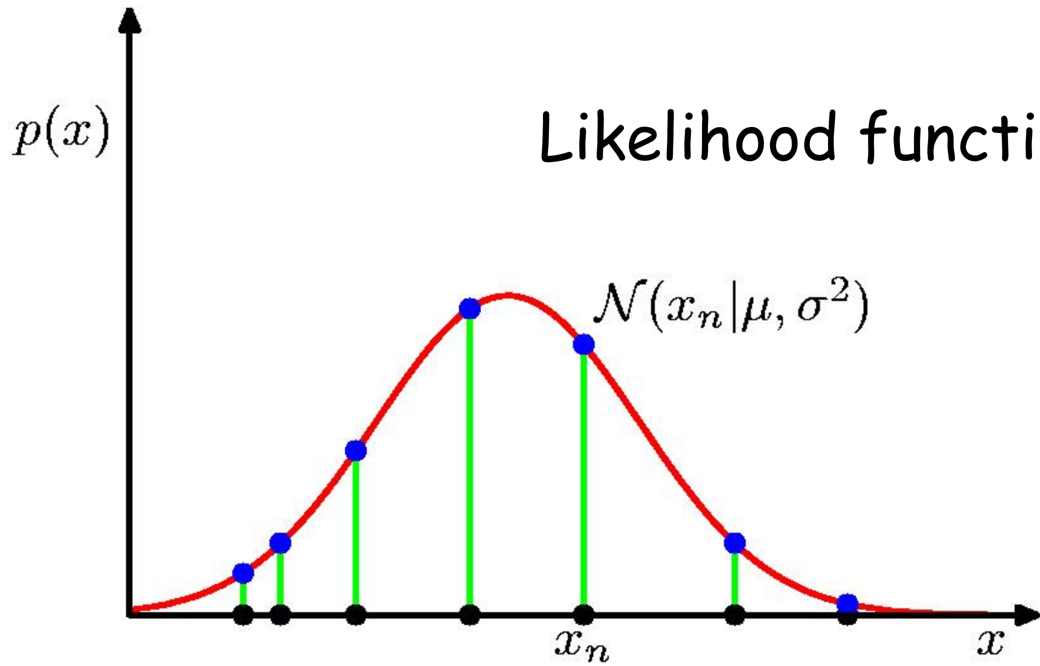
- Converting to the Standard Normal Distribution

$$z = \frac{x - \mu}{\sigma}$$

We can think of z as a measure of the number of standard deviations x is from μ.

Gaussian Parameter Estimation

Likelihood function



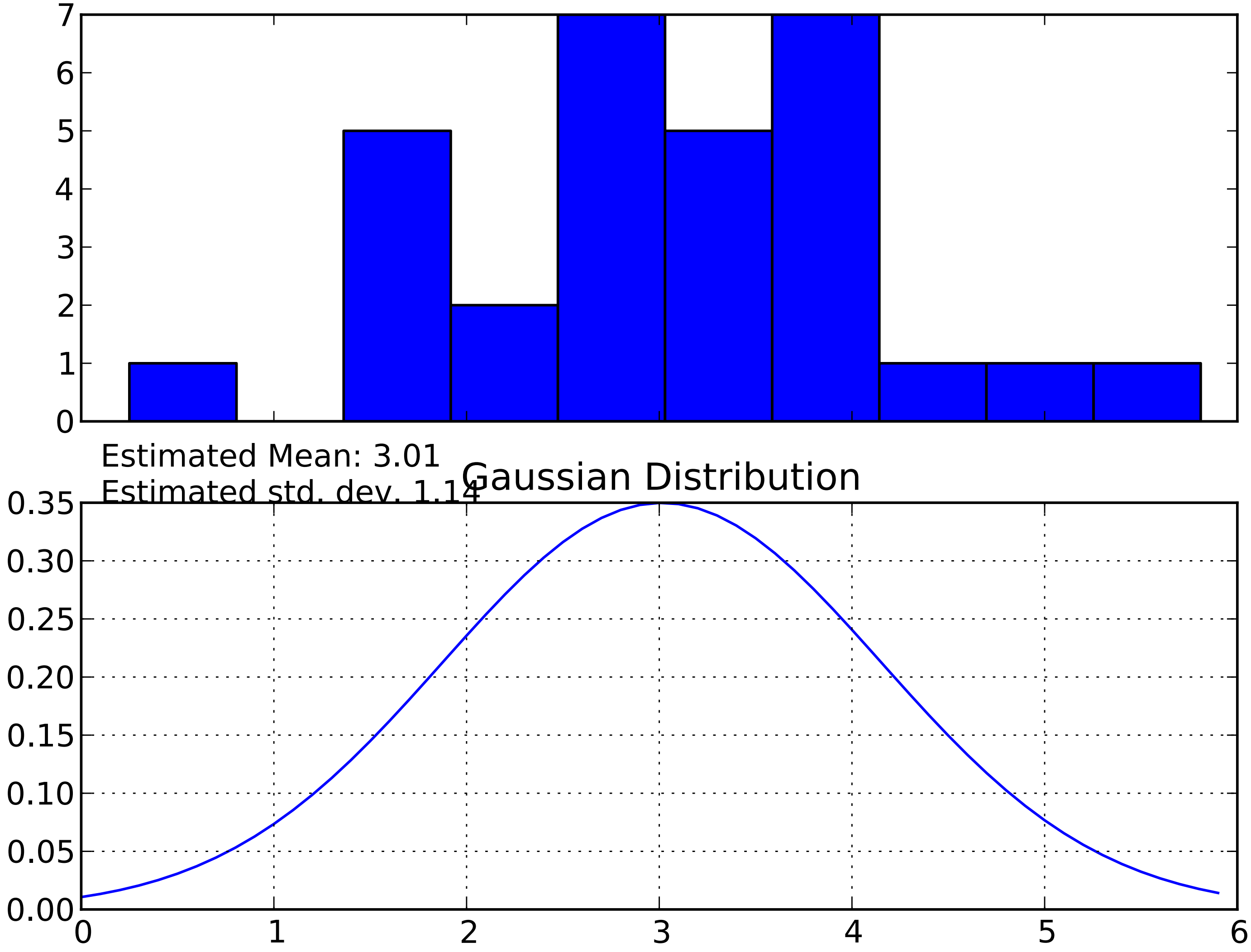
$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

Maximum (Log) Likelihood

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

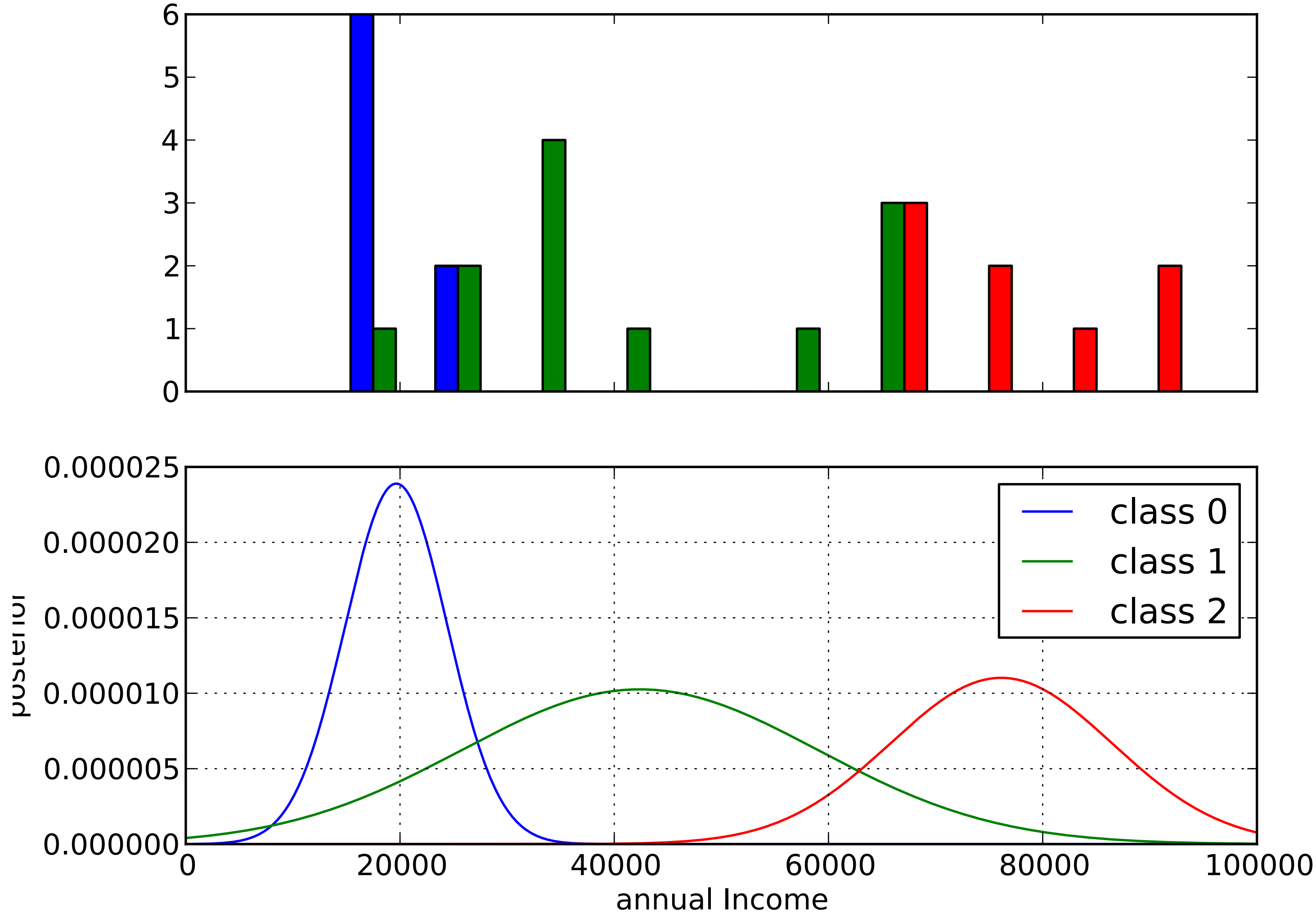
Example



Estimated Mean: 3.01
Estimated std. dev. 1.14
Gaussian Distribution
0
1
2
3
4
5
6
7
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35

Gaussian models

Assume we have data that belongs to three classes, and assume a likelihood that follows a Gaussian distribution



Gaussian Naïve Bayes

Likelihood function:

$$P(X_i = x|Y = y_k) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} \exp\left(-\frac{(x-\mu_{ik})^2}{2\sigma_{ik}^2}\right)$$

Need to estimate mean and variance for each feature in each class.

Summary

Naïve Bayes classifier:

- What's the assumption
- Why we make it
- How we learn it

Naïve Bayes for discrete data

Gaussian naïve Bayes