A few concepts from Probability Theory and Statistics

Alberto Paccanaro

EMAp – FGV

www.paccanarolab.org

Material and images in these slides are from (or adapted from):
C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006

Why probability ?

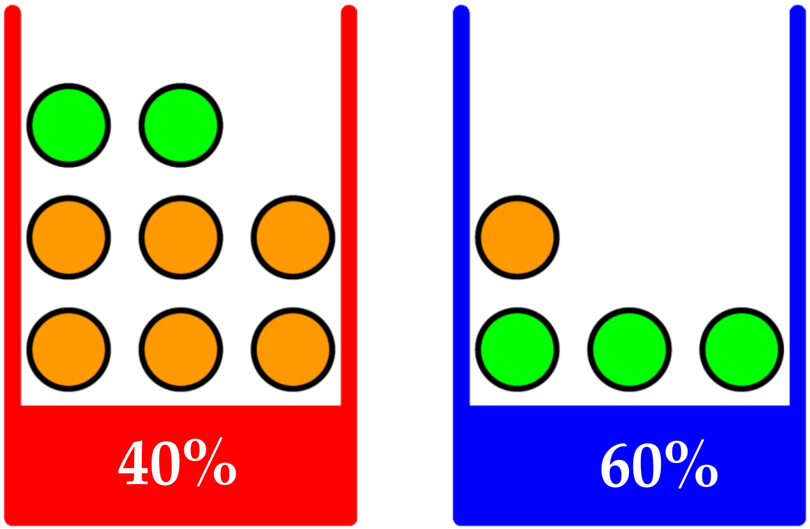
→ **Uncertainty**

Where does it come from?

1. noise on measurements
2. finite size of data sets

Probability theory provides a framework for handling uncertainty

Problem

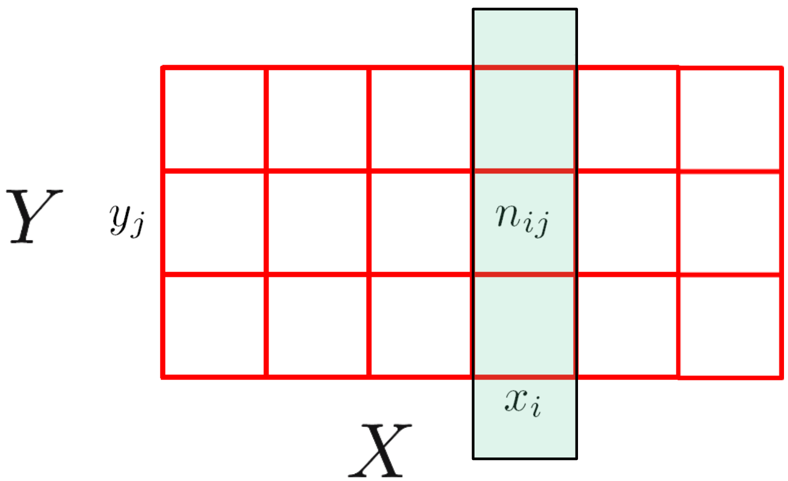


***Probability of an event**: fraction of times that event occurs out of the total number of trials (N), as N* $\rightarrow \infty$

Questions:

1. what is the overall probability to pick an apple ?
2. given that I picked an orange, what is the probability that I picked it from the red box?

Rules of Probability



$p(X = x_i\,, Y = y_j)$ indicates the probability that X will take the value x_i and Y will take the value y_j (***joint* probability**)

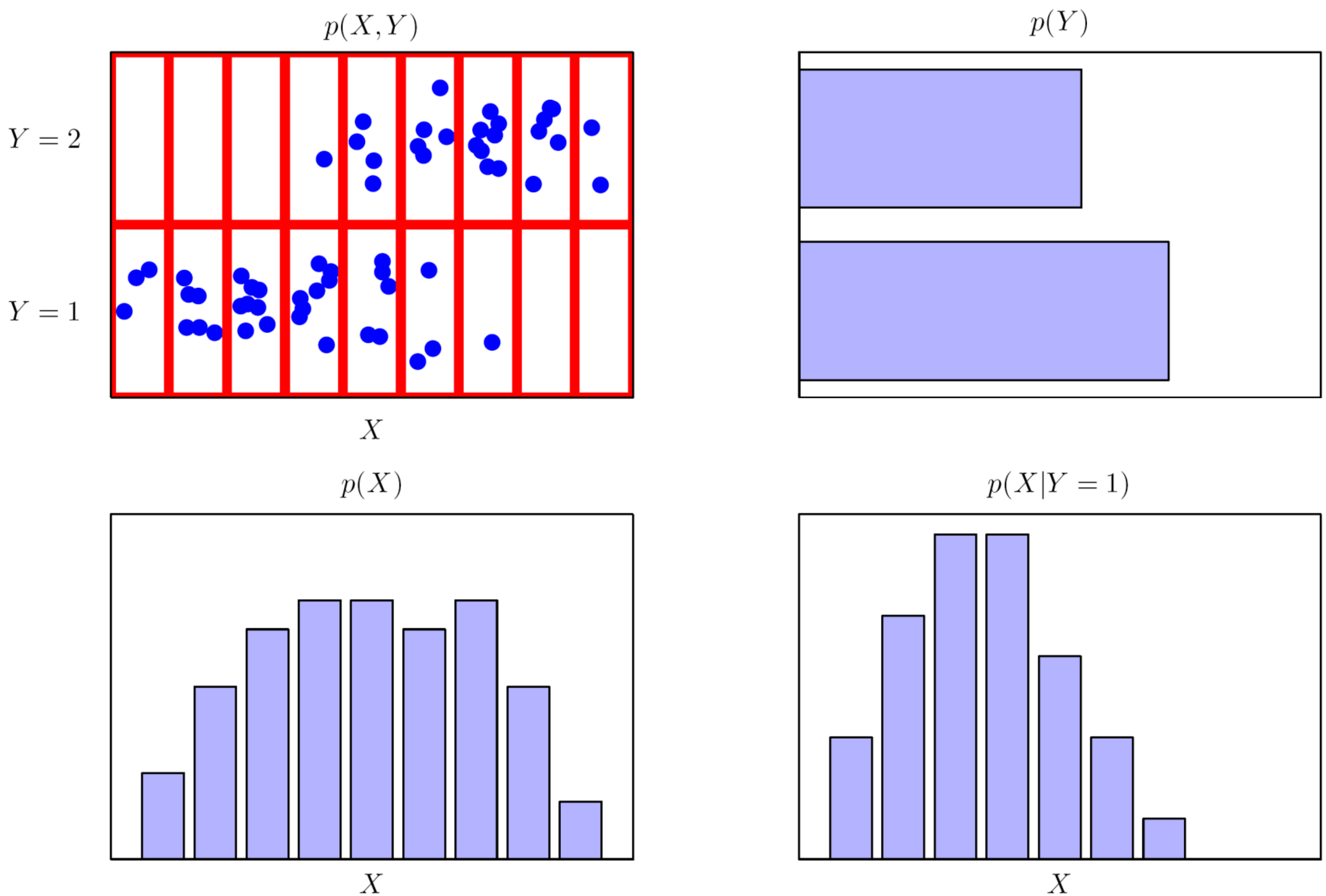
Consider instances for which $X = x_i$. $p(Y = y_j\ |X = x_i)$ is the fraction of such instances for which $Y = y_j$ (***conditional* probability** of Y given X).

sum rule $$p(X) = \sum_Y p(X, Y)$$

product rule $$p(X, Y) = p(Y|X)p(X)$$

If $p(Y|X) = p(Y)$ then $p(X, Y) = p(X)p(Y)$ and X and Y are said to be **independent**.

Another example



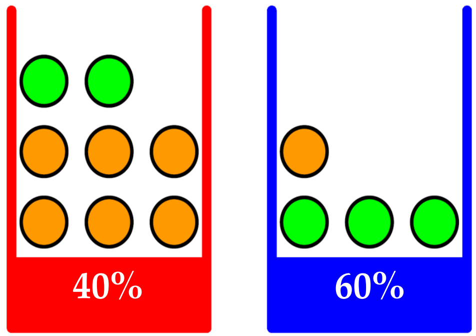
Bayes Theorem

$$\boxed{p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}}$$

The denominator in Bayes' theorem is a normalization constant:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}$$

… and back to the box of fruit



Two variables: F (fruit) can be a or o
B (box) can be r or b

Questions:

1. *what is the overall probability to pick an apple?*
2. *given that I picked an orange, what is the probability that I picked it from the red box?*

1. $P(F = a)$
2. $P(B = r \mid F = o)$

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

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

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

$$\begin{aligned} p(B=r) &= \\ p(B=b) &= \end{aligned} \quad ?$$

$$\begin{aligned} p(F=a|B=r) &= \\ p(F=o|B=r) &= \\ p(F=a|B=b) &= \\ p(F=o|B=b) &= \end{aligned} \quad ?$$

$$p(F=a) = p(F=a|B=r)p(B=r) + p(F=a|B=b)p(B=b)$$

9/20

$$p(B=r|F=o) = \frac{p(F=o|B=r)p(B=r)}{p(F=o)}$$

2/3

Bayes theorem, posterior probability, prior probability

Let's go back to the second question: *"given that I picked an orange, what is the probability that I picked it from the red box?"*

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

$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)}$$

Before being told the identity of the selected item of fruit, my best answer would have been $p(B)$ – the ***prior probability*** because it is the probability available ***before*** we observe the identity of the fruit.

Once I am told that the fruit is an orange, we can then use Bayes' theorem to compute the probability $p(B|F)$ – the ***posterior probability*** because it is the probability obtained ***after*** we have observed F

Probability densities

Probabilities with respect to continuous variables

Probability of a real-valued variable x falling in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta x \to 0$

$p(x)$ is the *probability density* over x

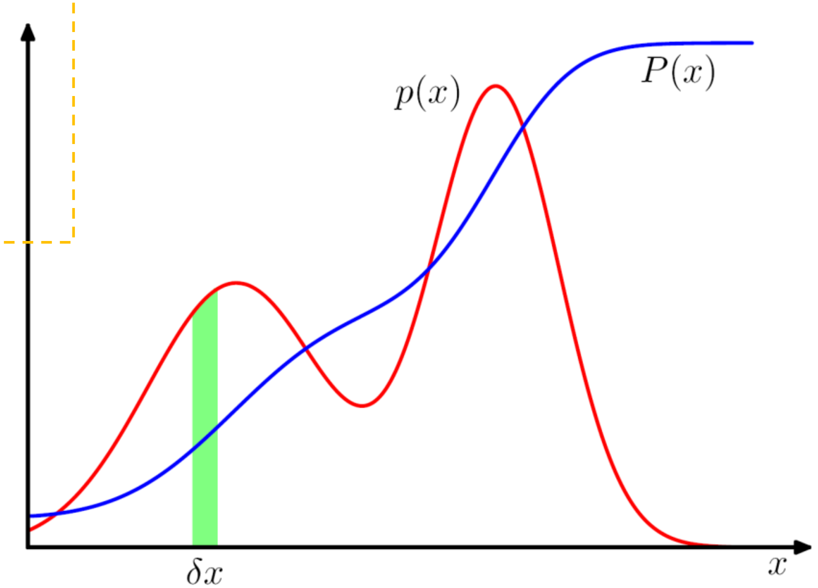
Properties:

$$p(x) \geqslant 0$$

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Cumulative distribution function (probability that x lies in the interval $(-\infty, z)$)

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$



Expectations and Covariances

Expectation of $f(x)$: average value of some function $f(x)$ under a probability distribution $p(x)$

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

Variance of $f(x)$: measures the variability of $f(x)$ around its mean.

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]$$

Covariance of two random variables x and y, expresses the extent to which x and y vary together

$$cov[x, y] = \mathbb{E}[\{\mathrm{x} - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}]$$

Covariance matrix of a random vector $x \in \mathrm{R}^n$ is an $n \times n$ matrix

$$\mathrm{Cov}(\mathbf{x})_{i,j} = \mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_j)$$
$$\mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_i) = \mathrm{Var}(\mathrm{x}_i)$$

Bayesian probabilities

We want to quantify the uncertainty that surrounds the the model parameters **w**

- Prior probability distribution $p(\mathbf{w})$: our assumptions about **w**, before observing the data
- Conditional probability $p(D|\mathbf{w})$: quantifies the effect of the data

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

- Posterior probability $p(\mathbf{w}|D)$: quantifies the uncertainty in **w** *after* we have observed D

$p(D \mid \mathbf{w})$: is evaluated for the observed data set D

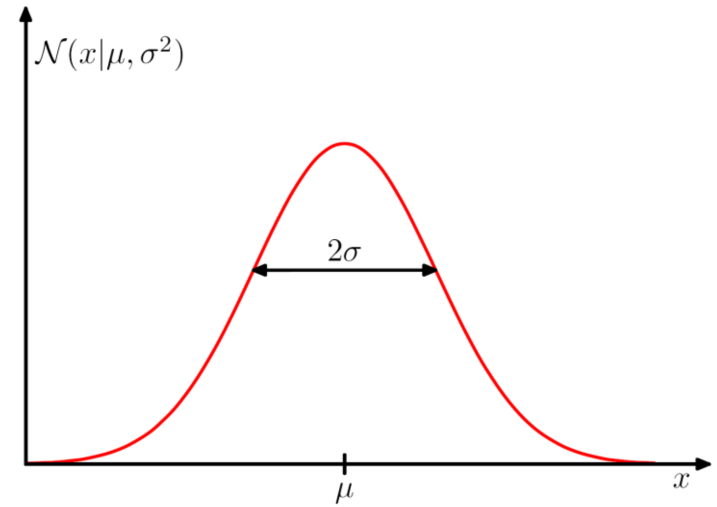
It is a function of the parameter vector **w**, the *lihelihood function*. It expresses how probable the observed data set is for different settings of the parameter vector **w**.

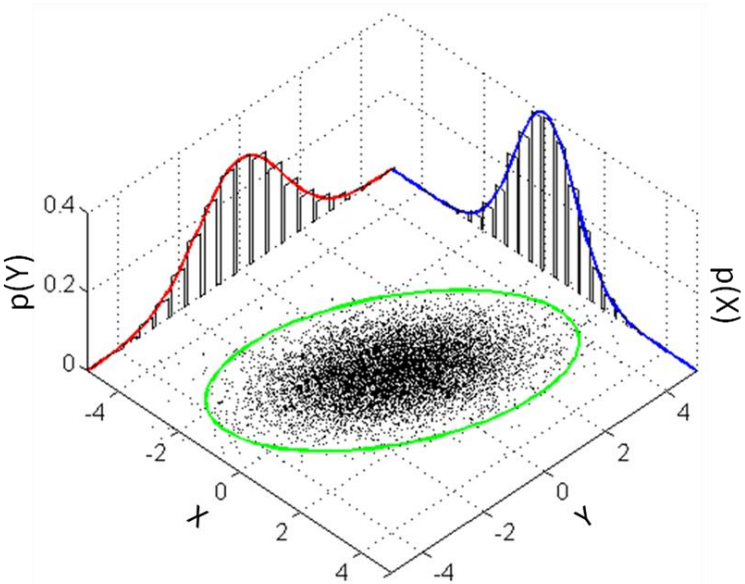
$$\boxed{\text{posterior} \propto \text{likelihood} \times \text{prior}}$$

"In a frequentist setting, **w** *is considered to be a fixed parameter, whose value is determined by some form of 'estimator', and error bars on this estimate are obtained by considering the distribution of possible data sets D.*

By contrast, from the Bayesian viewpoint there is only a single data set D (namely the one that is actually observed), and the uncertainty in the parameters is expressed through a probability distribution over **w***."*

The Gaussian distribution





from wikipedia

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

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Why, the Gaussian distribution?

(Under mild conditions) the sum of a set of random variables has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases