Lecture 3: Probabilistic Learning

DD2431

Giampiero Salvi, with contributions from A. Maki and J. Sullivan

Autumn, 2015

Probability vs Heuristics

Heuristic

experience-based techniques for problem solving, learning, and discovery that give a solution which is not guaranteed to be optimal (Wikipedia)

Typical examples:

- Artificial Neural Networks
- Decision Trees
- Evolutionary methods
- k-nearest neighbor

Discriminative vs Generative Models

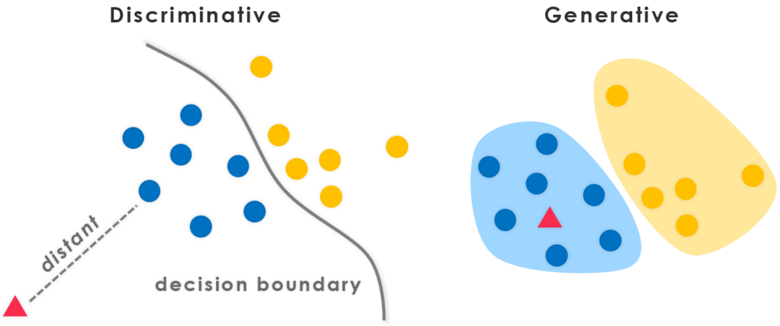
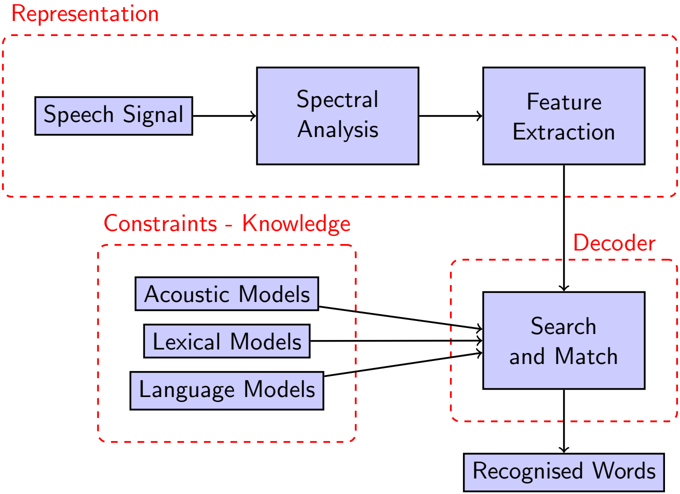


Figure from Nguyen *et al.* 2015. `http://www.evolvingai.org/fooling`

Advantages of Probability Based Methods

- **Results are interpretable.** More transparent and mathematically rigorous than methods such as *ANN*, *Evolutionary methods*.
- **Tool for interpreting other methods.** Framework for formalizing other methods - *concept learning*, *least squares*.
- **Work with sparse training data.** More powerful than deterministic methods when training data is sparse (framework for including prior knowledge).
- **Easy to merge different parts of a complex system.**

Example: Automatic Speech Recognition



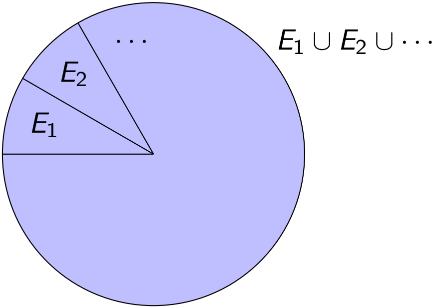
Different views on probabilities

Axiomatic	defines axioms and derives properties
Classical	number of ways something can happen over total number of things that can happen (e.g. dice)
Logical	same, but weight the different ways
Frequency	frequency of success in repeated experiments
Subjective	degree of belief (basis for Bayesian statistics)

Axiomatic definition of probabilities (Kolmogorov)

Given an event E in a event space F

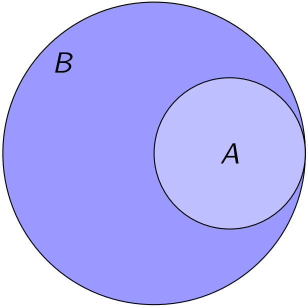
1. $P(E) \geq 0$ for all $E \in F$
2. sure event Ω: $P(\Omega) = 1$
3. $E_1, E_2, \ldots$ countable sequence of pairwise disjoint events, then



$$P(E_1 \cup E_2 \cup \cdots) = \sum_{i=1}^{\infty} P(E_i)$$

Consequences

1. Monotonicity: $P(A) \leq P(B)$ if $A \subseteq B$



Example: $A = \{3\}, \quad B = \{\text{odd}\}$

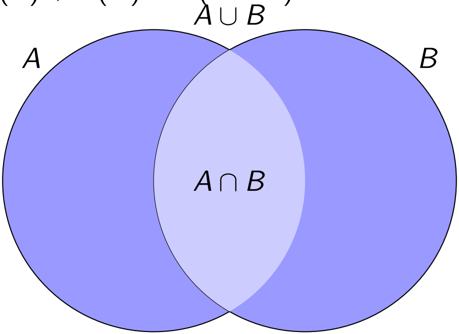
2. Empty set $\emptyset$: $P(\emptyset) = 0$

Example: $P(A \cap B)$ where $A = \{\text{odd}\}, B = \{\text{even}\}$

3. Bounds: $0 \leq P(E) \leq 1$ for all $E \in F$

More Consequences: Addition

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

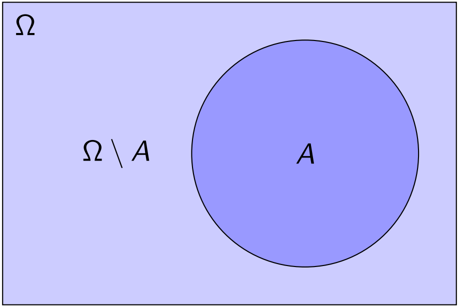


Example:

$$\begin{aligned} A &= \{1,3,5\}, & P(A) &= \frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{1}{2} \\ B &= \{5,6\}, & P(B) &= \frac{1}{6}+\frac{1}{6}=\frac{1}{3} \\ A \cap B &= \{5\} & P(A \cap B) &= \frac{1}{6} \\ A \cup B &= \{1,3,5,6\} & P(A \cup B) &= \frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{2}{3} \end{aligned}$$

More Consequences: Negation

$$P(\bar{A}) = P(\Omega \setminus A) = 1 - P(A)$$



Example:

$$\begin{aligned} A &= \{1,2\}, & P(A) &= \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{3} \\ \bar{A} &= \{3,4,5,6\}, & P(\bar{A}) &= \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = 1 - \tfrac{1}{3} \end{aligned}$$

Random (Stochastic) Variables

A random variable is a function that assigns a number x to the outcome of an experiment

- the result of flipping a coin,
- the result of measuring the temperature

The *probability distribution* $P(x)$ of a random variable (r.v.) captures the fact that

- the r.v. will have different values when observed **and**
- some values occur more than others.

Formal definition of RVs

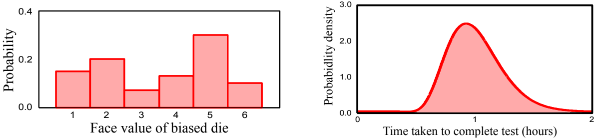
$$RV = \{f : \mathcal{S}_a \rightarrow \mathcal{S}_b, P(x)\}$$

where:

$$\begin{aligned}
\mathcal{S}_a &= \text{set of possible outcomes of the experiment} \\
\mathcal{S}_b &= \text{domain of the variable} \\
f : \mathcal{S}_a \rightarrow \mathcal{S}_b &= \text{function mapping outcomes to values } \; x \\
P(x) &= \text{probability distribution function}
\end{aligned}$$

Types of Random Variables

- A **discrete random variable** takes values from a predefined set.
- For a **Boolean discrete random variable** this predefined set has two members - $\{0, 1\}$, $\{\text{yes}, \text{no}\}$ etc.
- A **continuous random variable** takes values that are real numbers.



discrete pdf **continuous pdf**

Figures taken from **Computer Vision: models, learning and inference** by Simon Prince.

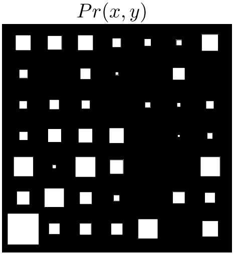
Examples of Random Variables

- Discrete events: either 1, 2, 3, 4, 5, or 6.
- Discrete probability distribution $p(x) = P(d = x)$
- $P(d = 1) = 1/6$ (fair dice)

- Any real number (theoretically infinite)
- Probability Distribution Function (PDF) $f(x)$ (NOT PROBABILITY!!!)
- $P(t = 36.6) = 0$
- $P(36.6 < t < 36.7) = 0.1$

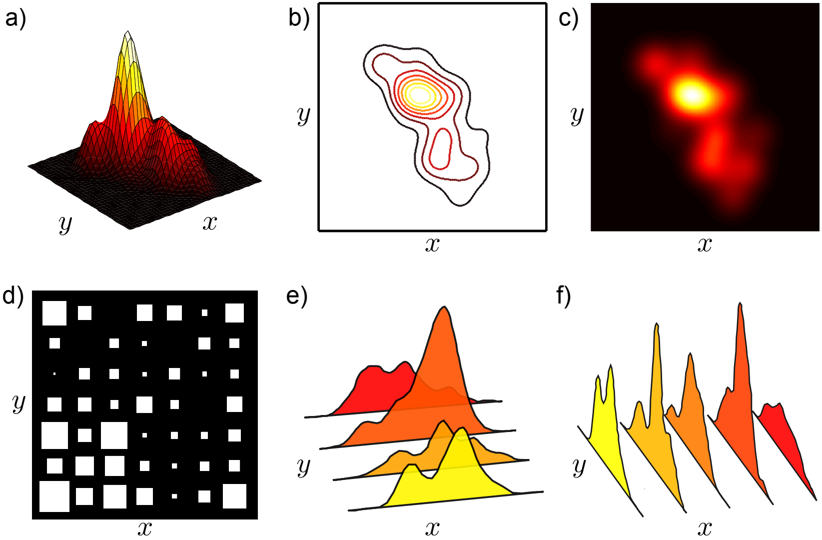
Joint Probabilities

- Consider two random variables x and y.
- Observe multiple paired instances of x and y. Some paired outcomes will occur more frequently.
- This information is encoded in the joint probability distribution $P(x, y)$.
- $P(\mathbf{x})$ denotes the joint probability of $\mathbf{x} = (x_1, \ldots, x_K)$.



$\leftarrow$ **discrete joint pdf**

Joint Probabilities (cont.)



Marginalization

The probability distribution of any single variable can be recovered from a joint distribution by summing for the discrete case

$$P(x) = \sum_y P(x, y)$$

and integrating for the continuous case

$$P(x) = \int_y P(x, y)\, dy$$

Marginalization (cont.)

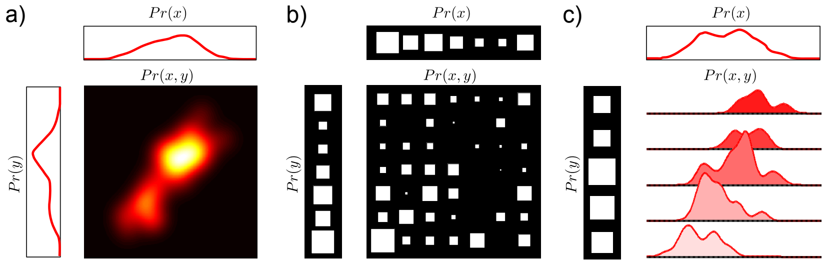


Figure from **Computer Vision: models, learning and inference** by Simon Prince.

Conditional Probabilities

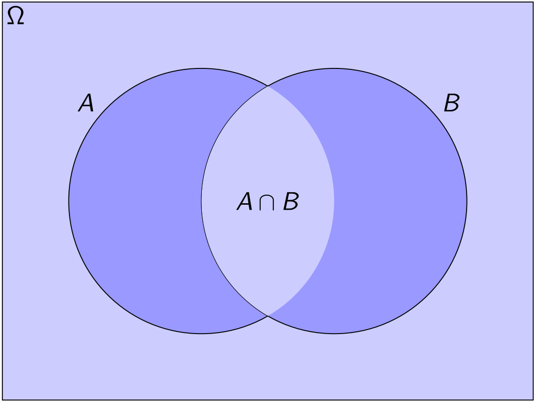
$$P(A|B)$$

The probability of event A when we *know* that event B has happened

Note: different from the probability that event A *and* event B will happen

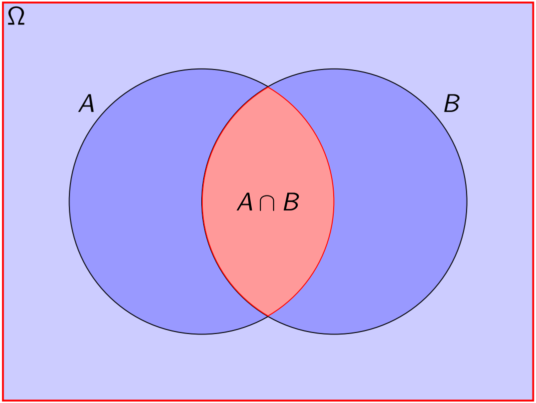
Conditional Probabilities

$$P(A|B) \neq P(A \cap B)$$



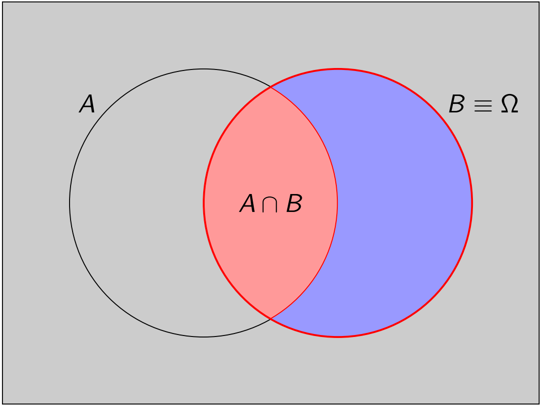
Conditional Probabilities

$$P(A|B) \neq P(A \cap B)$$



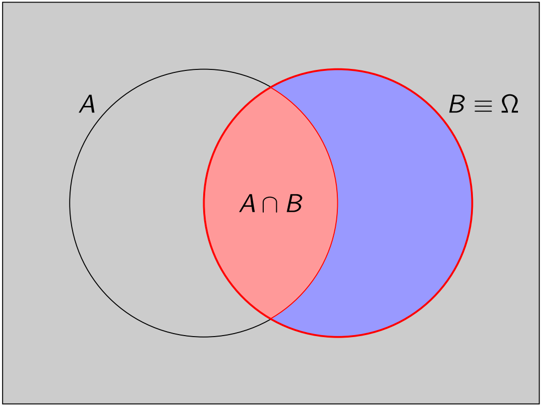
Conditional Probabilities

$$P(A|B) \neq P(A \cap B)$$



Conditional Probabilities

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

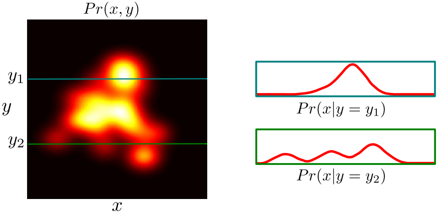


Conditional Probability (Random Variables)

- The conditional probability of x given that y takes value y^* indicates the different values of r.v. x which we'll observe given that y is fixed to value y^*.
- The conditional probability can be recovered from the joint distribution $P(x, y)$:

$$P(x \mid y = y^*) = \frac{P(x, y = y^*)}{P(y = y^*)} = \frac{P(x, y = y^*)}{\int_x P(x, y = y^*)\, dx}$$

- Extract an appropriate slice, and then normalize it.



Bayes' Rule

if

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

then

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

and

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

Bayes' Rule (random variables)

Bayes' Rule

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)} = \frac{P(x \mid y)P(y)}{\sum_y P(x \mid y)P(y)}$$

Each term in Bayes' rule has a name:

- $P(y \mid x) \leftarrow$ *Posterior* (what we know about y given x.)
- $P(y) \leftarrow$ *Prior* (what we know about y before we consider x.)
- $P(x \mid y) \leftarrow$ *Likelihood* (propensity for observing a certain value of x given a certain value of y)
- $P(x) \leftarrow$ *Evidence* (a constant to ensure that the l.h.s. is a valid distribution)

Independence

- two events are independent if the joint distribution can be factorized: $P(A \cap B) = P(A)P(B)$
- this means that:

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

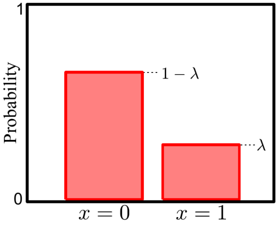
knowing that B has happened does not tell us anything about A

Bernoulli

- Domain: binary variables ($x \in \{0, 1\}$)
- Parameters: $\lambda = Pr(x = 1)$, $\lambda \in [0, 1]$

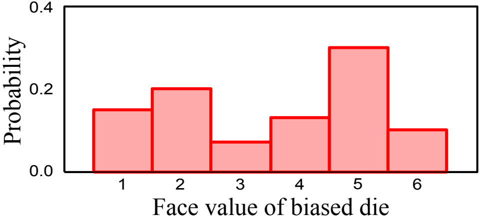
Then $Pr(x = 0) = 1 - \lambda$, and

$$Pr(x) = \lambda^x (1 - \lambda)^{1-x} = \begin{cases} \lambda, & \text{if } x = 1, \\ 1 - \lambda, & \text{if } x = 0 \end{cases}$$



Categorical

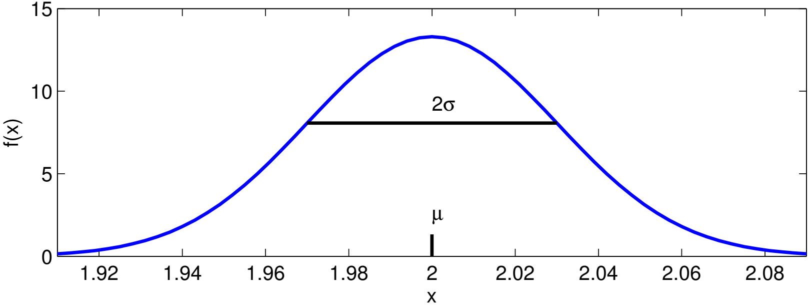
- Domain: discrete variables ($x \in \{x_1, \dots, x_K\}$)
- Parameters: $\lambda = [\lambda_1, \dots, \lambda_K]$
- with $\lambda_k \in [0, 1]$ and $\sum_{k=1}^{K} \lambda_k = 1$



Gaussian distributions: One-dimensional

- aka univariate normal distribution
- Domain: real numbers ($x \in \mathbb{R}$)

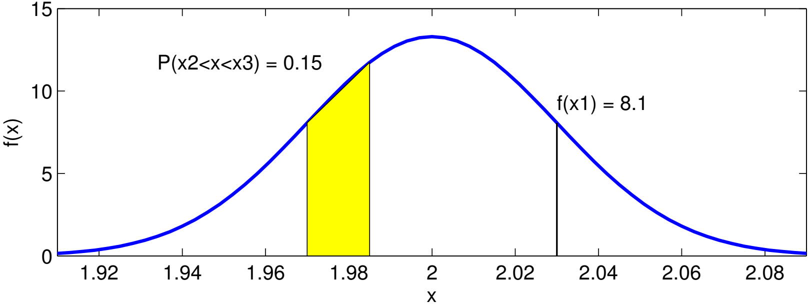
$$f(x|\mu,\sigma^2) = N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



Gaussian distributions: One-dimensional

- aka univariate normal distribution
- Domain: real numbers ($x \in \mathbb{R}$)

$$f(x|\mu,\sigma^2) = N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



Gaussian distributions: D Dimensions

- aka multivariate normal distribution
- Domain: real numbers ($\mathbf{x} \in \mathbb{R}^D$)

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_D \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_D \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1D} \\ \sigma_{21} & \dots & & \\ \dots & & & \\ \sigma_{D1} & \dots & & \sigma_{DD}^2 \end{bmatrix}$$

$$f(\mathbf{x}|\mu, \Sigma) = \frac{exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right]}{(2\pi)^{\frac{D}{2}}\,|\Sigma|^{\frac{1}{2}}}$$

Covariance and Independence

- covariance is "linear" dependency
- dependent variables may have zero covariance
- in Gaussian (and few other distribution) zero covariance is equivalent to independence

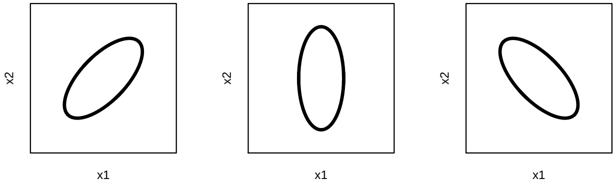
$$f(\mathbf{x}|\mu, \Sigma) = \frac{exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right]}{(2\pi)^{\frac{D}{2}}\,|\Sigma|^{\frac{1}{2}}}$$

Gaussian distributions

$$f(\mathbf{x}|\mu,\Sigma) = \frac{exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right]}{(2\pi)^{\frac{D}{2}}\,|\Sigma|^{\frac{1}{2}}}$$

Eigenvalue decomposition of the covariance matrix:

$$\Sigma = \lambda\; R\; \Sigma_{\text{diag}}\; R^T$$



Beta and Dirichlet (PDF over Probabilities)

Beta

- Domain: real numbers, bounded ($\lambda \in [0, 1]$)
- Parameters: $\alpha, \beta \in \mathbb{R}_+$
- describes probability of parameter λ in Bernoulli

Dirichlet

- Domain: K real numbers, bounded ($\lambda_1, \ldots, \lambda_K \in [0, 1]$)
- Parameters: $\alpha_1, \ldots, \alpha_K \in \mathbb{R}_+$
- describes probability of parameters λ_k in Categorical

General ML problem (supervised learning)

Data:

$$\{(\mathbf{x}^1, y^1), (\mathbf{x}^2, y^2), \ldots, (\mathbf{x}^n, y^n)\}$$

Where $\mathbf{x}$ are features, and y is the answer

- if y is discrete: classification
- if y is continuous: regression

- Learning: we observe several examples of $\mathbf{x}$ and we know y
- Inference: we want to know y given a new $\mathbf{x}$

Machine Learning with Probabilities

Learning: we observe several examples of $\mathbf{x}$ and we know y

- we can estimate $P(y)$ and $P(\mathbf{x}|y)$

Inference: we want to know y given a new $\mathbf{x}$

- we want to estimate $P(y|\mathbf{x})$

- $P(\mathbf{x} \mid y) \leftarrow$ *Likelihood* represents the probability of observing data $\mathbf{x}$ given the hypothesis y.
- $P(y) \leftarrow$ *Prior of y* represents the background knowledge of hypothesis y being correct.
- $P(y \mid \mathbf{x}) \leftarrow$ *Posterior* represents the probability that hypothesis y is true after data $\mathbf{x}$ has been observed.

Bayes' Rule

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

With

- $P(\mathbf{x} \mid y) \leftarrow$ *Likelihood* represents the probability of observing data $\mathbf{x}$ given the hypothesis y.
- $P(y) \leftarrow$ *Prior of y* represents the background knowledge of hypothesis y being correct.
- $P(y \mid \mathbf{x}) \leftarrow$ *Posterior* represents the probability that hypothesis y is true after data $\mathbf{x}$ has been observed.

Learning and Inference

- **Bayesian Learning:** The process of learning the likelihood distribution $P(\mathbf{x} \mid y)$ and prior probability distribution $P(y)$ from a set of training points

$$\{(\mathbf{x}^1, y^1), (\mathbf{x}^2, y^2), \ldots, (\mathbf{x}^n, y^n)\}$$

- **Bayesian Inference:** The process of calculating the posterior probability distribution $P(y \mid \mathbf{x})$ for certain data $\mathbf{x}$.

Example: Which Gender?

Task: Determine the gender of a person given their measured hair length.

Notation:

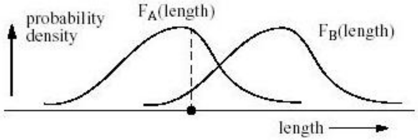
- Let $g \in \{\text{'f'}, \text{'m'}\}$ be a r.v. denoting the gender of a person.
- Let x be the measured length of the hair.

Information given:

- The hair length observation was made at a boy's school thus

$$P(g = \text{'m'}) = .95, \quad P(g = \text{'f'}) = .05$$

- Knowledge of the likelihood distributions $P(x \,|\, g = \text{'f'})$ and $P(x \,|\, g = \text{'m'})$



Example: Which Gender?

Task: Determine the gender of a person given their measured hair length $\implies$ calculate $P(g \,|\, x)$.

Solution:

Apply Bayes' Rule to get

$$P(g = \text{'m'} \,|\, x) = \frac{P(x \,|\, g = \text{'m'})P(g = \text{'m'})}{P(x)}$$
$$= \frac{P(x \,|\, g = \text{'m'})P(g = \text{'m'})}{P(x \,|\, g = \text{'f'})P(g = \text{'f'}) + P(x \,|\, g = \text{'m'})P(g = \text{'m'})}$$

Can calculate $P(g = \text{'f'} \,|\, x) = 1 - P(g = \text{'m'} \,|\, x)$

Selecting the most probably hypothesis

- **Maximum A Posteriori (*MAP*) Estimate**:

  Hypothesis with highest probability given observed data

$$\begin{aligned} y_{\mathrm{MAP}} &= \arg\max_{y \in \mathcal{Y}} P(y \mid \mathbf{x}) \\ &= \arg\max_{y \in \mathcal{Y}} \frac{P(\mathbf{x} \mid y)\, P(y)}{P(\mathbf{x})} \\ &= \arg\max_{y \in \mathcal{Y}} P(\mathbf{x} \mid y)\, P(y) \end{aligned}$$

- **Maximum Likelihood Estimate (*MLE*):**

  Hypothesis with highest likelihood of generating observed data.

$$y_{\mathrm{MLE}} = \arg\max_{y \in \mathcal{Y}} P(\mathbf{x} \mid y)$$

  Useful if we do not know prior distribution or if it is uniform.

Example: Cancer or Not?

Scenario:
A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, 0.8% of the entire population have cancer.

Scenario in probabilities:

- **Priors:**

$$P(\text{disease}) = .008 \qquad P(\text{not disease}) = .992$$

- **Likelihoods:**

$$P(+\,|\,\text{disease}) = .98 \qquad P(+\,|\,\text{not disease}) = .03$$
$$P(-\,|\,\text{disease}) = .02 \qquad P(-\,|\,\text{not disease}) = .97$$

Example: Cancer or Not?

Find MAP estimate:
When test returned a positive result,

$$\begin{aligned} y_{\text{MAP}} &= \arg \max_{y \in \{\text{disease, not disease}\}} P(y \,|\, +) \\ &= \arg \max_{y \in \{\text{disease, not disease}\}} P(+ \,|\, y)\, P(y) \end{aligned}$$

Substituting in the correct values get

$$\begin{aligned} P(+ \,|\, \text{disease})\, P(\text{disease}) &= .98 \times .008 = .0078 \\ P(+ \,|\, \text{not disease})\, P(\text{not disease}) &= .03 \times .992 = .0298 \end{aligned}$$

Therefore $y_{\text{MAP}} =$ "not disease".

The Posterior probabilities:

$$\begin{aligned} P(\text{disease} \,|\, +) &= \frac{.0078}{(.0078 + .0298)} = .21 \\ P(\text{not disease} \,|\, +) &= \frac{.0298}{(.0078 + .0298)} = .79 \end{aligned}$$