Probability, Decision and Information Theory

TTT4185 Machine Learning for Signal Processing

Giampiero Salvi

Department of Electronic Systems
NTNU

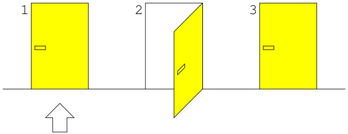
HT2021

Probability Theory in ML

incorporate probabilistic thinking at all levels

- start with incomplete knowledge (uncertainty)
- use observations to reduce uncertainty
- belief propagation

probability distributions as carriers of information[1]



[1] E T Jaynes. *Probability theory: The logic of science*. Ed. by G Larry Bretthorst. Cambridge university press, June 2003.

Engineering vs Science

Engineering:

- ML as collection of methods
- fine tune aspects to boost the results

Science:

- define unified theory
- give the deepest possible interpretation to the results

Engineering vs Science

Engineering:

- ML as collection of methods
- fine tune aspects to boost the results

Science:

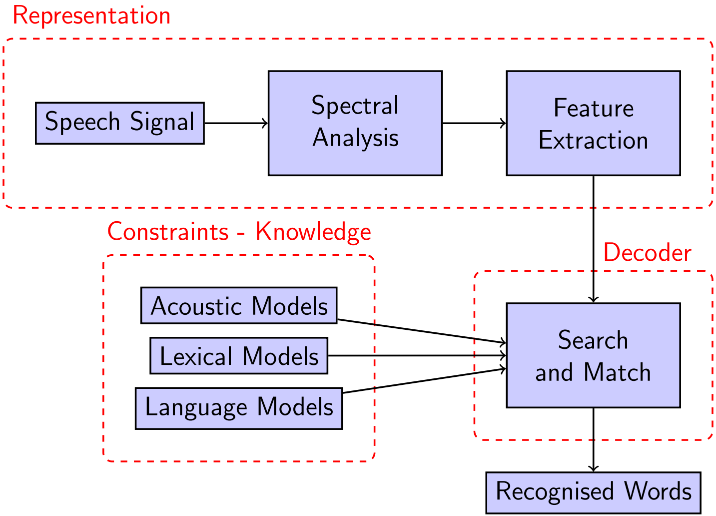
- define unified theory
- give the deepest possible interpretation to the results

reality not 100% clear cut

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.** And make the assumptions explicit — *concept learning*, *least squares*.
- **Work with sparse training data.** More powerful than deterministic methods when training data is sparse (framework for including prior knowledge).
- **Belief Propagation:** Easy to merge different parts of a complex system and to update current knowledge with new observations.

Example: Automatic Speech Recognition



Advantages of Probability Based Methods, ctnd.

- **Shape a way of thinking.** All aspects of learning, modelling and inference can be cast under the same theory.

Disadvantages of Probability Based Methods

- **Often hard to derive closed solutions.** Need to resort to heuristic approximations.
- **Inefficient for large data sets.** But many argue that the need for large data set is a flaw in the methods.

Outline

1 Probability Theory Reminder
- Axioms and Properties
- Common Distributions
- Moments

2 Probabilistic Machine Learning
- Supervised Learning, General Definition
- Regression
- Classification
- Bayes decision theory

3 Information Theory

Outline

1. Probability Theory Reminder
   - Axioms and Properties
   - Common Distributions
   - Moments
2. Probabilistic Machine Learning
   - Supervised Learning, General Definition
   - Regression
   - Classification
   - Bayes decision theory
3. Information Theory

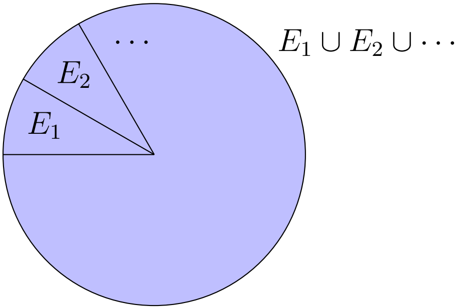
Different views on probabilities

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

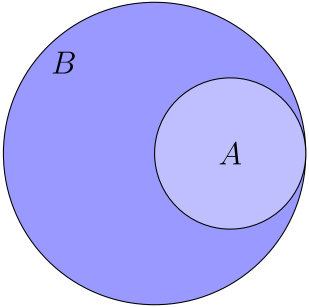
1. $P(E) \geq 0$ for all $E \in F$
2. sure event Ω: $P(\Omega) = 1$
3. $E_1, E_2, \ldots$ countable sequence of pairwise disjoint events, then



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

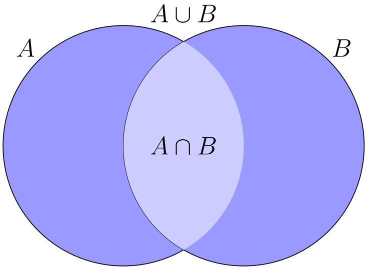
Consequences

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

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

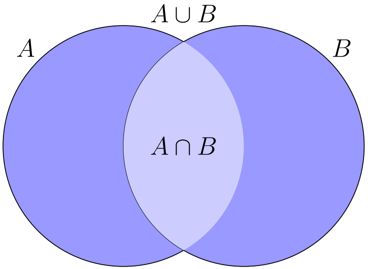


Example:

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

More Consequences: Addition

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

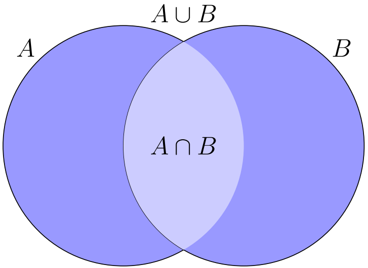


Example:

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

More Consequences: Addition

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

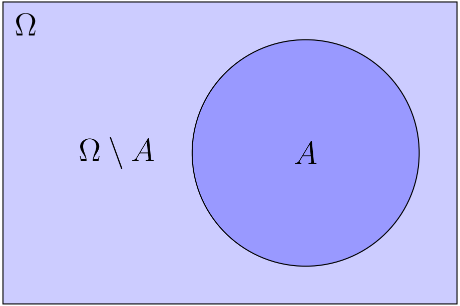


Example:

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

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

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

where:

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

Examples of RVs

Dice:

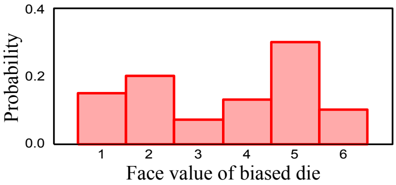
$$
\begin{aligned}
\mathcal{S}_a &= \text{the dice lands on one of the sides} \\
\mathcal{S}_b &= \text{integer numbers } \{1, 2, 3, 4, 5, 6\} \\
f : \mathcal{S}_a \to \mathcal{S}_b &= \text{assigns each side to a number} \\
P(x) &= \text{uniform distribution for a fair dice}
\end{aligned}
$$

Temperature:

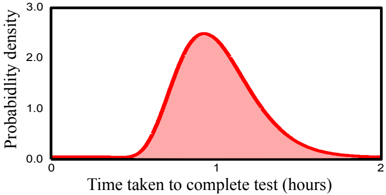
$$
\begin{aligned}
\mathcal{S}_a &= \text{degrees of expansion of mercury} \\
\mathcal{S}_b &= \text{real numbers } \mathbb{R} \\
f : \mathcal{S}_a \to \mathcal{S}_b &= \text{maps expansion to a real number (different maps} \\
& \quad \text{for Celsius, Fahrenheit or absolute temperature)} \\
P(x) &= \text{depends on the application}
\end{aligned}
$$

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

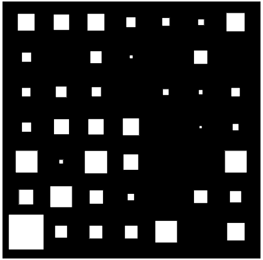
- Discrete events: either 1, 2, 3, 4, 5, or 6.
- Discrete probability distribution $p(x) = P(d = x)$
- $P(d = 1) = 1/6$ (fair dice)

- Any real number (theoretically infinite)
- Probability Density 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 $\mathbf{x} = (x_1, \ldots, x_K)$.

$Pr(x, y)$



$\leftarrow$ **discrete joint pdf**

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

Joint Probabilities (cont.)

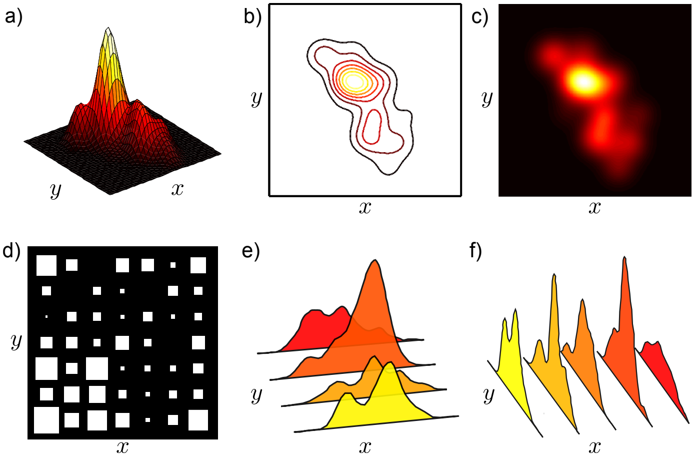


Figure from **Computer Vision: models, learning and inference** by Simon Prince.

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.)

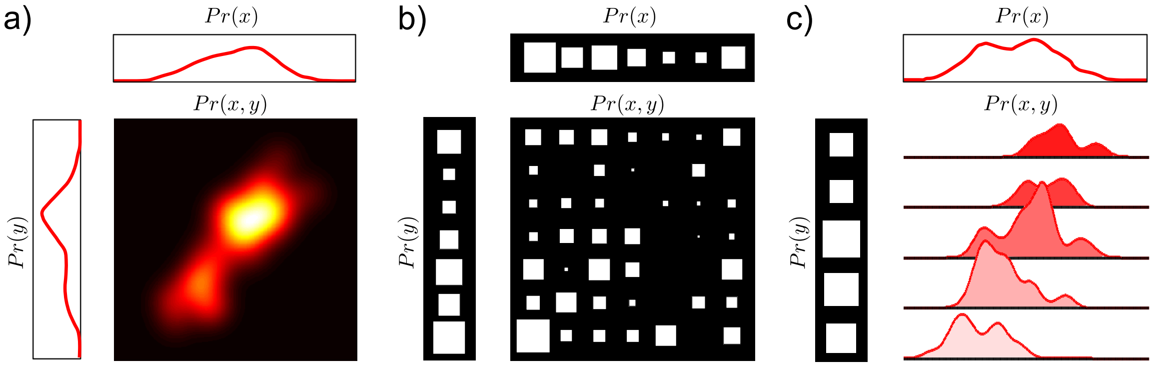


Figure from **Computer Vision: models, learning and inference** by Simon Prince.

Conditional Probabilities

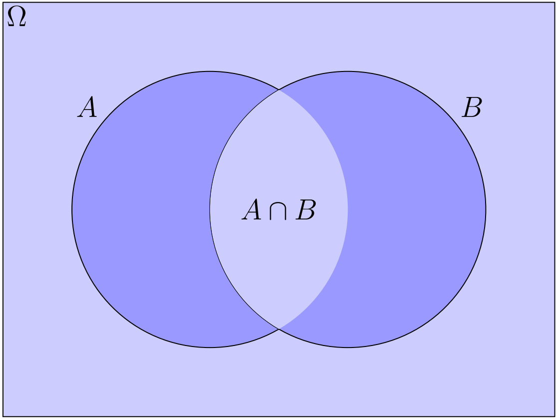
$$P(A|B)$$

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

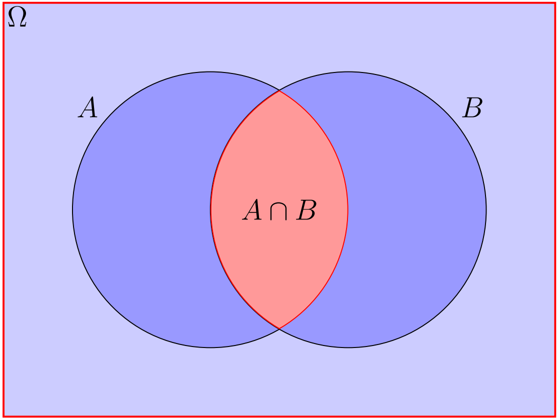
Conditional Probabilities

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



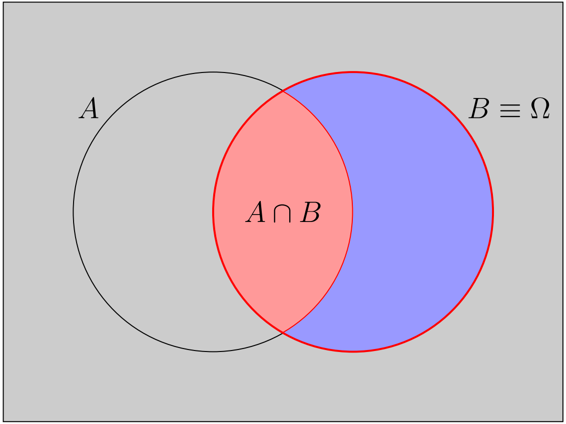
Conditional Probabilities

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



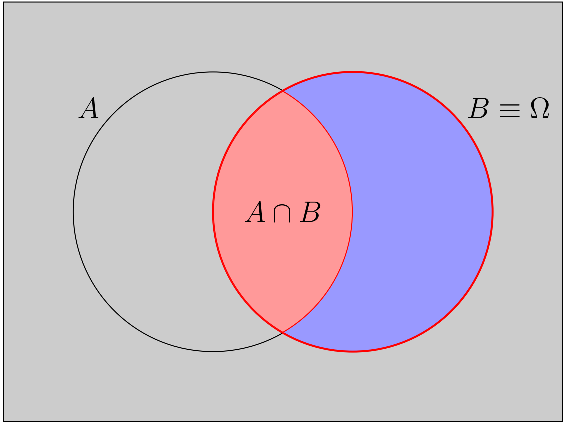
Conditional Probabilities

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



Conditional Probabilities

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

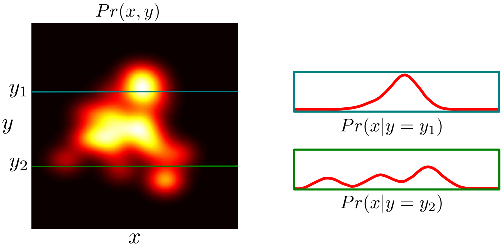


Conditional Probability (Random Variables)

- 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 $P(x, y)$:

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

- Extract an appropriate slice, and then normalize it.



Independence

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

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

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

Bayes' Rule

if

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

then

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

Bayes' Rule

if

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

then

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

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

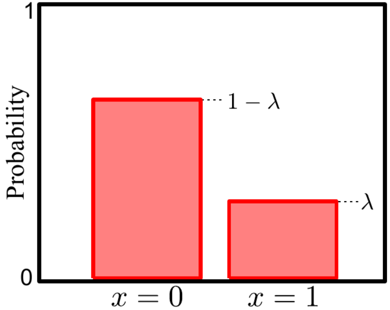
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bernoulli: binary variables

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

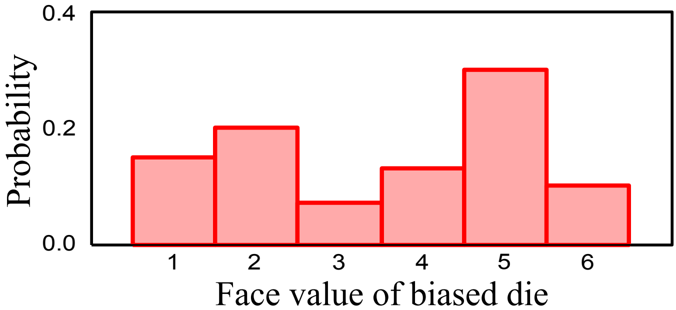
Then $Pr(x = 0) = 1 - \lambda$, and

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



Categorical

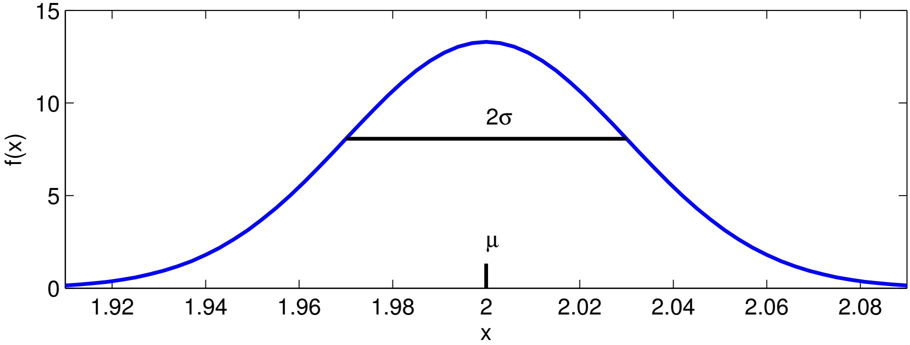
- Domain: discrete variables ($x \in \{x_1, \dots, x_K\}$)
- Parameters: $\lambda = [\lambda_1, \dots, \lambda_K]$
- with $\lambda_k \in [0, 1]$ and $\sum_{k=1}^{K} \lambda_k = 1$



Gaussian distributions: One-dimensional

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

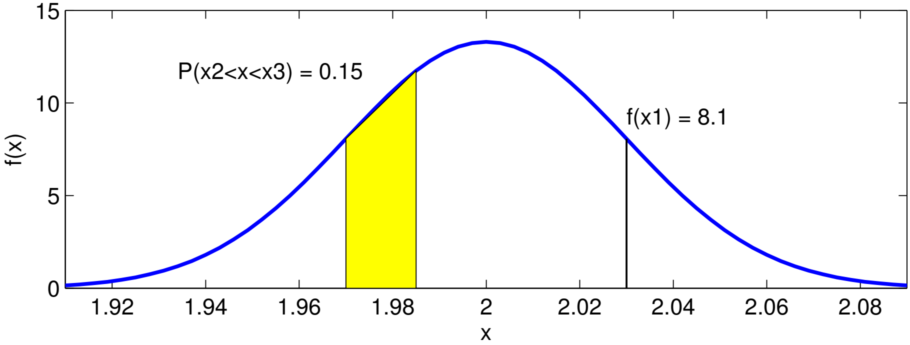
$$f(x|\mu,\sigma^2) = N(\mu,\sigma^2) = N(\mu,\beta^{-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



Gaussian distributions: One-dimensional

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

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

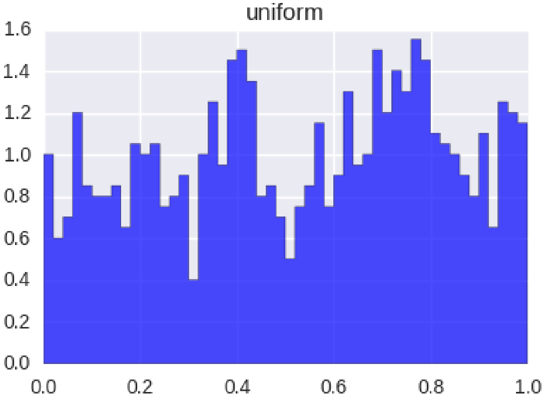


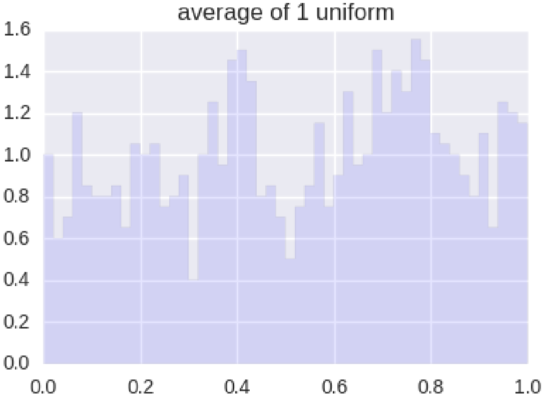
Why Gaussian: Central Limit Theorem

Galton Board (Sir Francis Galton, 1822-1911)

Why Gaussian: Central Limit Theorem

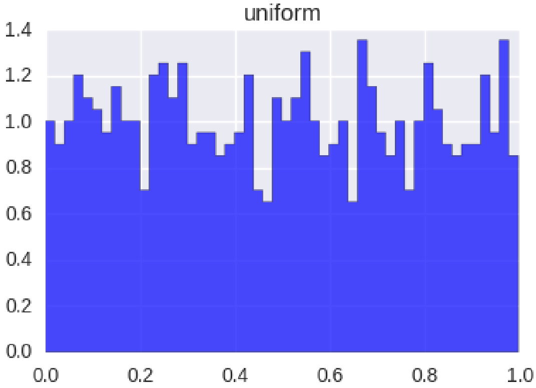
The distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.[2].

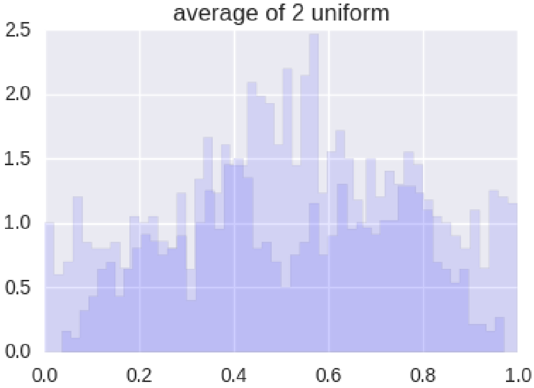




Why Gaussian: Central Limit Theorem

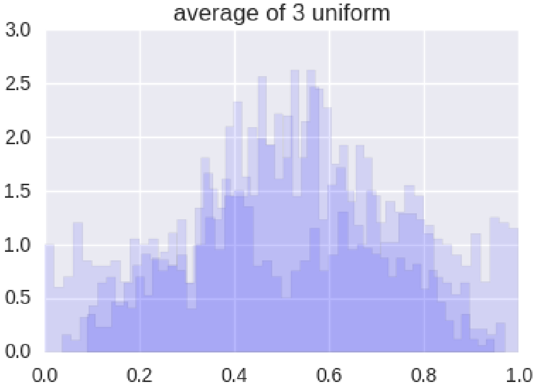
The distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.[2].





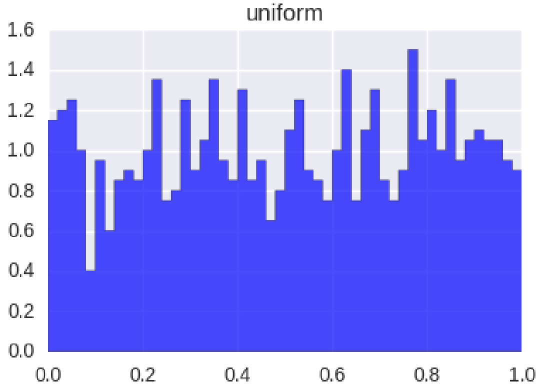
Why Gaussian: Central Limit Theorem

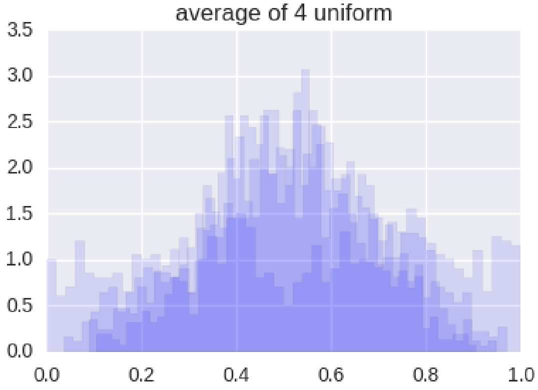
The distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.[2].



Why Gaussian: Central Limit Theorem

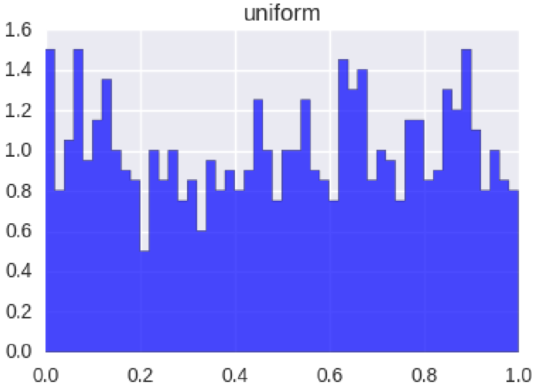
The distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.[2].

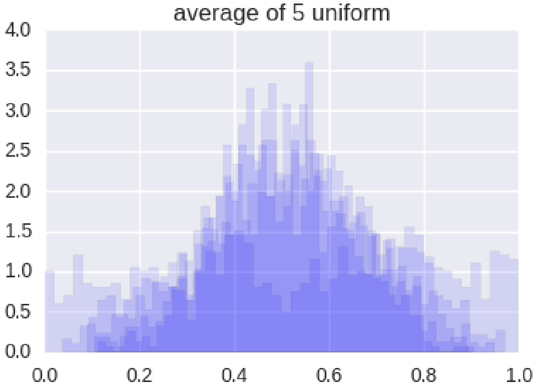




Why Gaussian: Central Limit Theorem

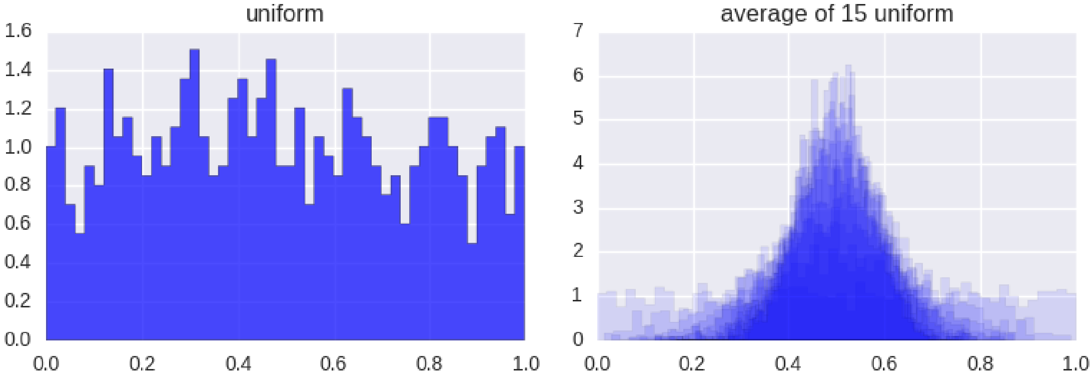
The distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.[2].





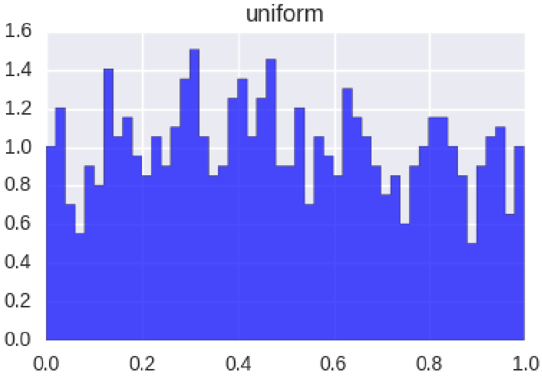
Why Gaussian: Central Limit Theorem

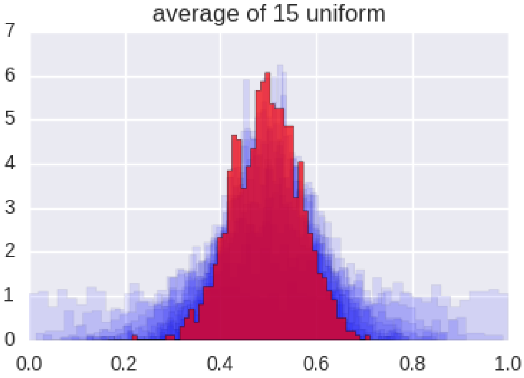
The distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.[2].



Why Gaussian: Central Limit Theorem

The distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.[2].





Gaussian distributions: D Dimensions

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

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_D \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_D \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1D} \\ \sigma_{21} & \sigma_2^2 & \dots & \\ \dots & & & \\ \sigma_{D1} & \dots & & \sigma_D^2 \end{bmatrix}$$

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

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:

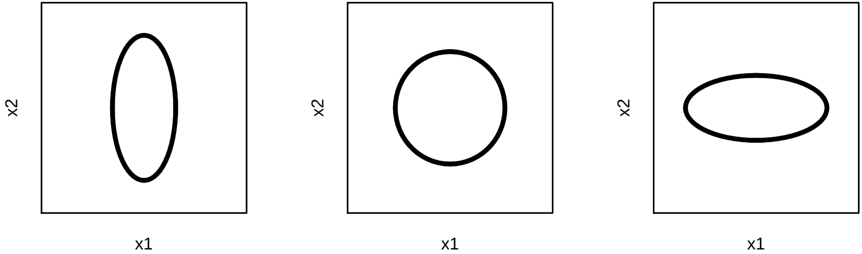
$$\Sigma = \lambda\; R\; \Sigma_{\text{diag}}\; R^T$$

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:

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

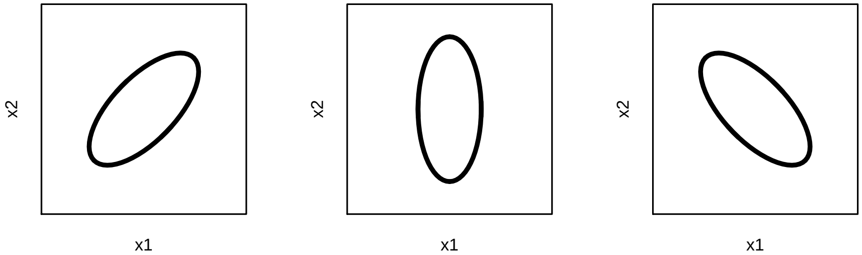


Gaussian distributions

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

Eigenvalue decomposition of the covariance matrix:

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

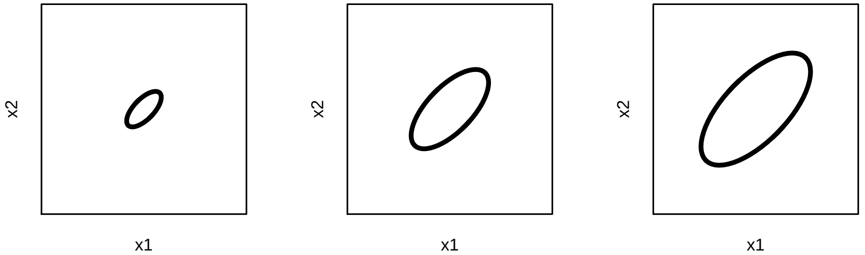


Gaussian distributions

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

Eigenvalue decomposition of the covariance matrix:

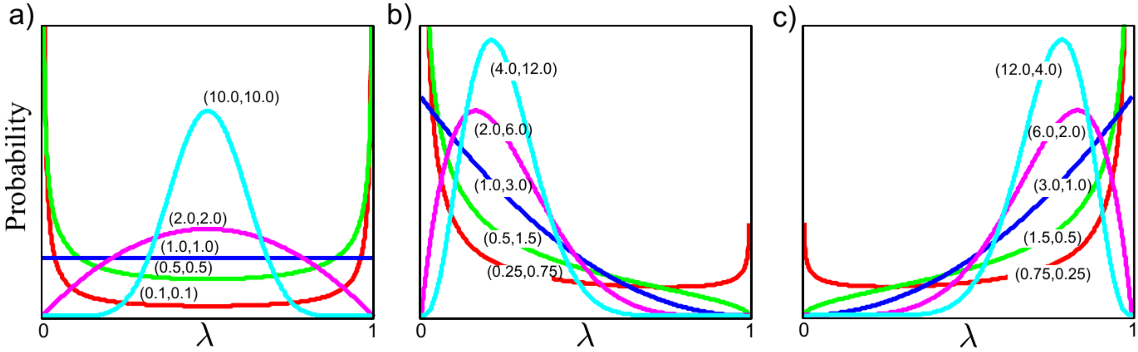
$$\Sigma = \lambda \; R \; \Sigma_{\text{diag}} \; R^T$$



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



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

Expected value

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

- Shows the "center of gravity" of a distribution
- Sampled expected value (mean)

$$\overline{\mathbf{x}} = \frac{1}{N} \sum_{i}^{N} \mathbf{x}_i$$

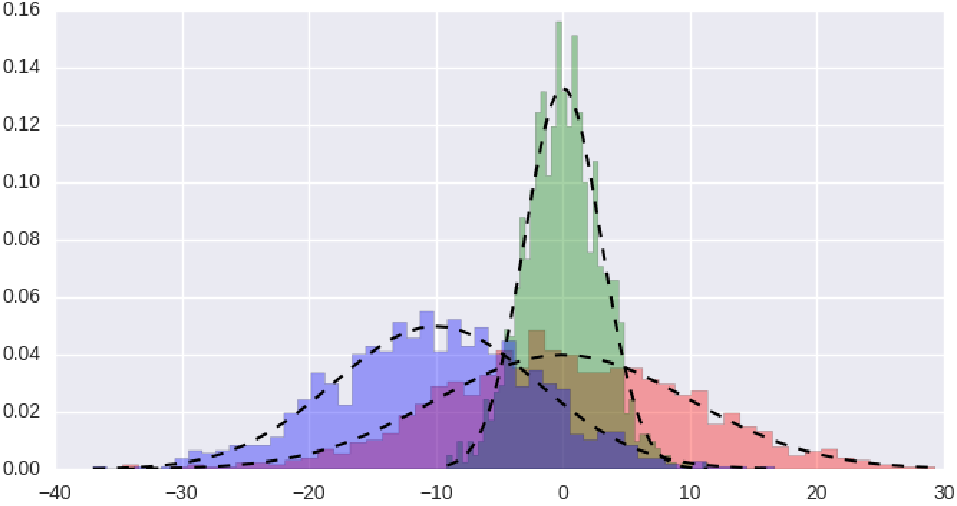
Variance

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

- Shows the “spread” of a distribution
- Sample variance

$$\overline{\sigma^2(\mathbf{x})} = \frac{1}{N-1}\sum_{i}^{N}(\mathbf{x}_i - \mu(\mathbf{x}_i))^2$$

Examples



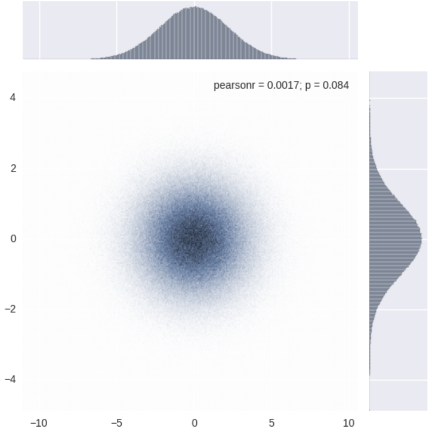
Covariance

$$\sigma(\mathbf{x}, \mathbf{y}) = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])]$$

- Shows how the “spread” of how to variables vary *together*
- Sample co-variance

$$\overline{\sigma(\mathbf{x}, \mathbf{y})} = \frac{1}{N-1} \sum_{i}^{N} (\mathbf{x}_i - \mu(\mathbf{x}_i))(\mathbf{y}_i - \mu(\mathbf{y}))$$

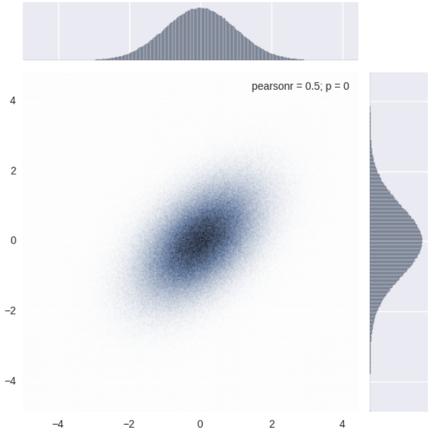
Examples



4

[4]Script by C.E. Ek

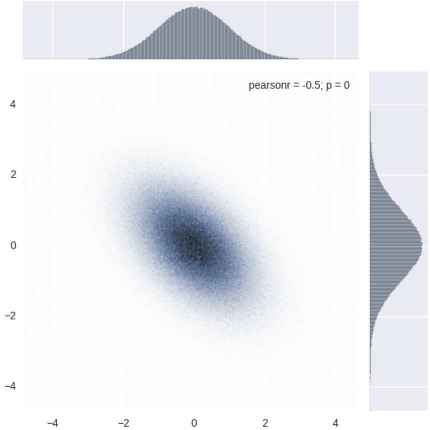
Examples



4

[4] Script by C.E. Ek

Examples

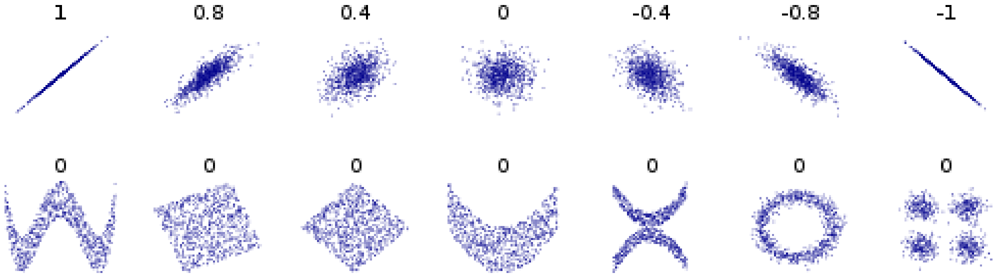


4

[4] Script by C.E. Ek

Covariance and Independence

- covariance is "linear" dependency
- dependent variables may have zero covariance
- in some distributions zero covariance is equivalent to independence



[5] Figure adapted from Wikipedia

Covariance and Independence (Gaussian)

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

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

Outline

1. Probability Theory Reminder
   - Axioms and Properties
   - Common Distributions
   - Moments
2. Probabilistic Machine Learning
   - Supervised Learning, General Definition
   - Regression
   - Classification
   - Bayes decision theory
3. Information Theory

General ML problem (supervised learning)

Data:

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

Where $\mathbf{x}$ are features, and t is the answer (target)

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

General ML problem (supervised learning)

Data:

$$\{(\mathbf{x}^1, t^1), (\mathbf{x}^2, t^2), \dots, (\mathbf{x}^n, t^n)\}$$

Where $\mathbf{x}$ are features, and t is the answer (target)

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

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

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

Inference: we want to know t' given a new $\mathbf{x}'$

- we want to estimate $P(t'|\mathbf{x}')$

Bayes' Rule

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

- $P(\mathbf{x} \mid t) \leftarrow$ Likelihood represents the probability of observing data $\mathbf{x}$ given the hypothesis t.
- $P(t) \leftarrow$ Prior represents the knowledge on hypothesis t before any observation.
- $P(t \mid \mathbf{x}) \leftarrow$ Posterior represents the probability of hypothesis t after the data $\mathbf{x}$ has been observed.
- $P(\mathbf{x}) \leftarrow$ Evidence encodes the quality of the underlying model.

$$P(\mathbf{x}) = \begin{cases} \sum_t P(\mathbf{x} \mid t)P(t) & \text{classification} \\ \int_t P(\mathbf{x} \mid t)P(t) & \text{regression} \end{cases}$$

Probabilistic Regression

Regression as conditional probability

- use multivariate joint distribution between $\mathbf{x}$ and t
- find posterior $p(t|\mathbf{x}, \theta)$ of t by conditioning on $\mathbf{x}$

Explicit regression model:

- define a deterministic model $t = y(\mathbf{x}, \theta) + \epsilon$
- describe probability distribution of error ϵ

Implicit model: Gaussian Processes

Example: bivariate Normal distribution

Define joint probability distribution function

$$\text{pdf}(x,t) = \mathcal{N}(x,t|\mu,\Sigma)$$

Where:

$$\mu = \begin{bmatrix} \mu_x \\ \mu_t \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_t \\ \rho\sigma_x\sigma_t & \sigma_t^2 \end{bmatrix}$$

and ρ is the correlation coefficient
Conditional probability distribution function still Normal:
$\text{pdf}(t|x,\mu,\Sigma) = \mathcal{N}(t, \mu_{t|x}, \sigma^2_{t|x})$, with:

$$\mu_{t|x} = \mu_t + \rho\frac{\sigma_t}{\sigma_x}(x - \mu_x) = w_0 + w_1 x$$
$$\sigma^2_{t|x} = (1-\rho^2)\sigma_t^2 \quad \text{(constant wrt. } x\text{)}$$

Explicit Regression Model

Model (deterministic):

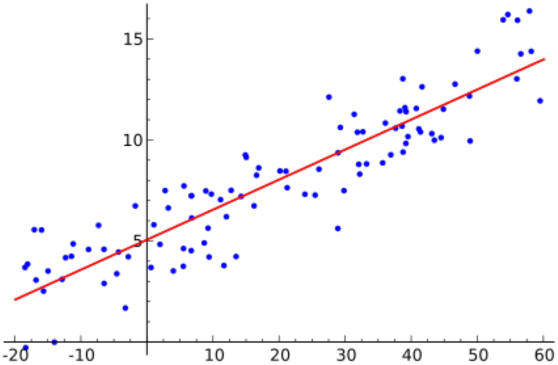
$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon$$

But now:

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

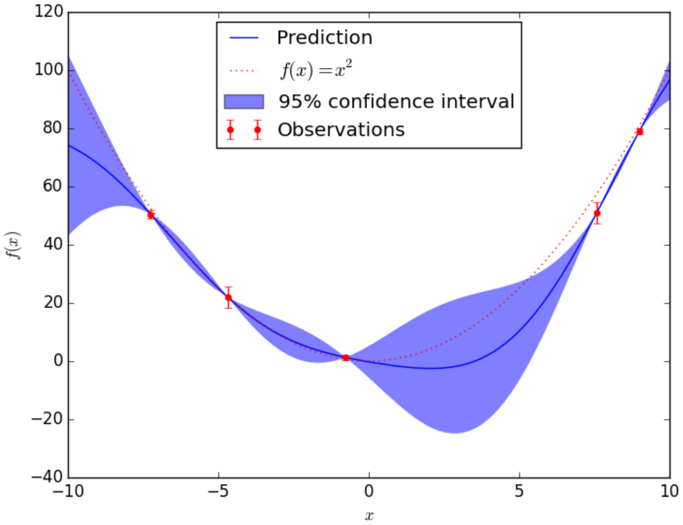
Therefore:

$$\begin{aligned} t &\sim \mathcal{N}(\mu_T(\mathbf{x}), \sigma_T^2(\mathbf{x})) \\ &= \mathcal{N}(y(\mathbf{x}, \mathbf{w}), \sigma^2) \end{aligned}$$



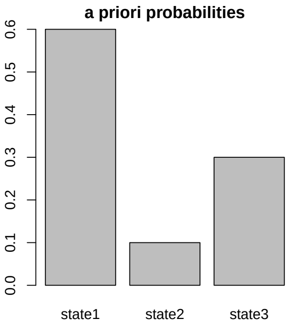
Gaussian Processes (advanced topic)

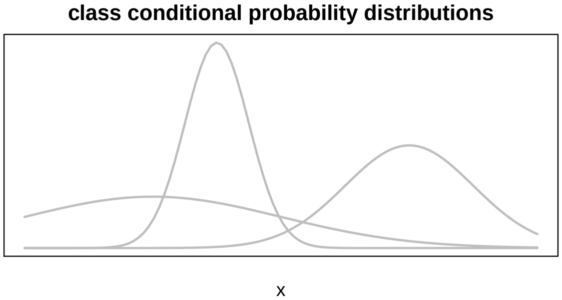
- non-parametric model
- covariance between t and $\mathbf{x}$ depends on observed data $\mathcal{D}$



The Probabilistic Model of Classification

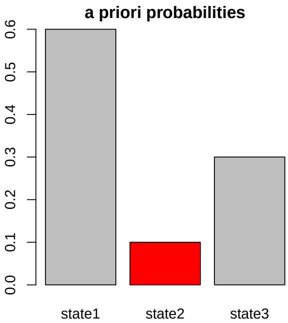
- one of k states t_j is selected with *a priori* probability $P(t_j)$
- When in state t_j, some observations $\hat{\mathbf{x}}$ are generated with distribution $p(\mathbf{x}|t_j)$

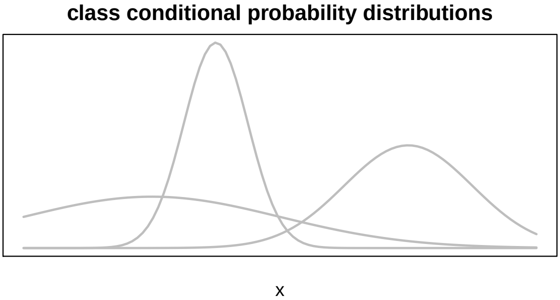




The Probabilistic Model of Classification

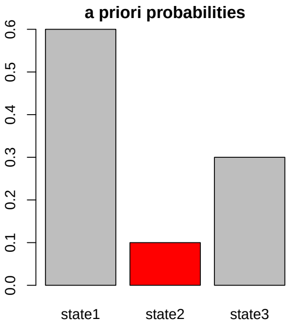
- one of k states t_j is selected with *a priori* probability $P(t_j)$
- When in state t_j, some observations $\hat{\mathbf{x}}$ are generated with distribution $p(\mathbf{x}|t_j)$

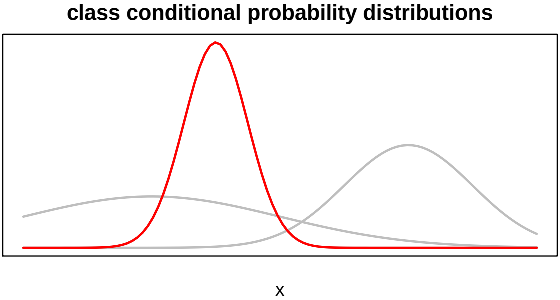




The Probabilistic Model of Classification

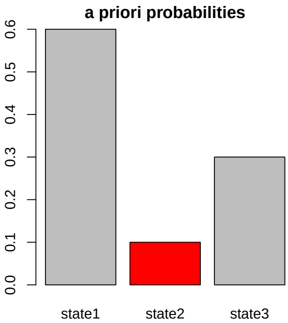
- one of k states t_j is selected with *a priori* probability $P(t_j)$
- When in state t_j, some observations $\hat{\mathbf{x}}$ are generated with distribution $p(\mathbf{x}|t_j)$

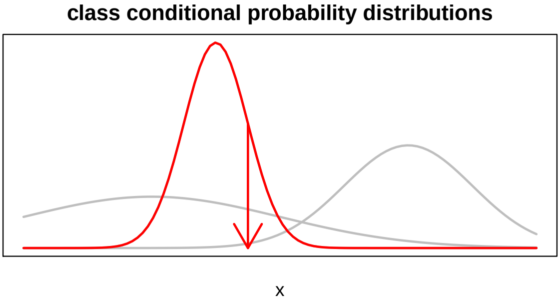




The Probabilistic Model of Classification

- one of k states t_j is selected with *a priori* probability $P(t_j)$
- When in state t_j, some observations $\hat{\mathbf{x}}$ are generated with distribution $p(\mathbf{x}|t_j)$





Problem

- If I observe a new $\hat{\mathbf{x}}$
- and I know $P(t_j)$ and $p(\mathbf{x}|t_j)$ for each class t_j
- what can I say about the state of the problem (class) t_j?
- equivalent to divideing feature space in K regions $\{\mathcal{R}_1, \ldots, \mathcal{R}_K\}$

Minimizing probability of errors

Example two classes:

$$
\begin{aligned}
P(\text{error}) &= P(\mathbf{x} \in \mathcal{R}_1, c_2) + P(\mathbf{x} \in \mathcal{R}_2, c_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, c_2) dx + \int_{\mathcal{R}_2} p(\mathbf{x}, c_1) d\mathbf{x}
\end{aligned}
$$

Minimizing probability of errors

Example two classes:

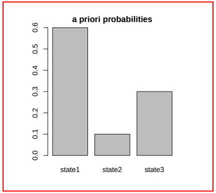
$$
\begin{aligned}
P(\text{error}) &= P(\mathbf{x} \in \mathcal{R}_1, c_2) + P(\mathbf{x} \in \mathcal{R}_2, c_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, c_2) dx + \int_{\mathcal{R}_2} p(\mathbf{x}, c_1) d\mathbf{x}
\end{aligned}
$$

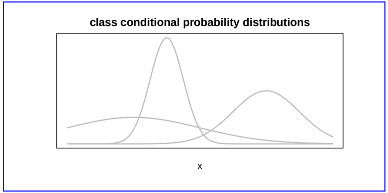
Example K classes (maximize correct):

$$
P(\text{correct}) = \sum_{k=1}^{K} P(\mathbf{x} \in \mathcal{R}_k, c_k) = \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, c_k) d\mathbf{x}
$$

In both cases equivalent to Maximum a posteriori (MAP)

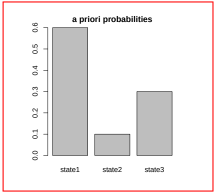
Bayes decision theory

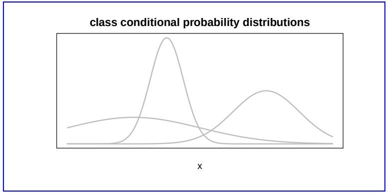




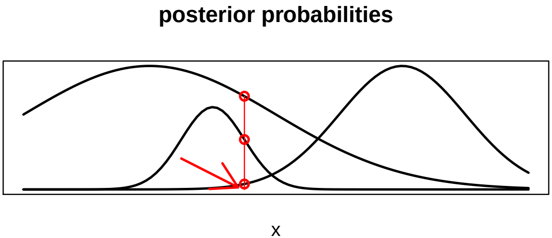
$$P(t_j|\hat{\mathbf{x}}) = \frac{p(\hat{\mathbf{x}}|t_j)\; P(t_j)}{p(\hat{\mathbf{x}})}$$

Bayes decision theory

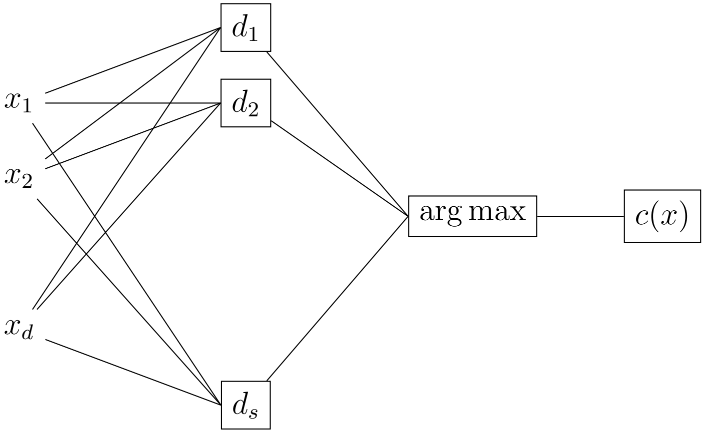




$$P(t_j|\hat{\mathbf{x}}) = \frac{p(\hat{\mathbf{x}}|t_j)\ P(t_j)}{p(\hat{\mathbf{x}})}$$



Classifiers: Discriminant Functions



$$d_i(\mathbf{x}) = p(\mathbf{x}|t_i)\ P(t_i)$$

Loss Function (classification)

$$L_{kj}, \quad k = \text{true class}, \quad j = \text{classification}$$

- different consequences for different kinds of errors:

Minimize expected loss

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(x, c_k) dx$$

Loss Function (regression)

$$L(t, y(\mathbf{x})), \quad t = \text{true value}, \quad y(\mathbf{x}) = \text{predicted value}$$

Expected loss:

$$\mathbb{E}[L] = \int\int L(t, y(\mathbf{x}))p(\mathbf{x}, t)d\mathbf{x}dt$$

Examples

$$\begin{aligned} L(t, y(\mathbf{x})) &= (y(\mathbf{x}) - t)^2 && \text{square loss} \\ L(t, y(\mathbf{x})) &= |y(\mathbf{x}) - t|^q && \text{Minkowski loss} \end{aligned}$$

$q = 2 \rightarrow$ conditional mean, $q = 1 \rightarrow$ conditional median, $q = 0 \rightarrow$ conditional mode

Example: Which Gender?

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

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.

Example: Which Gender?

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

Notation:

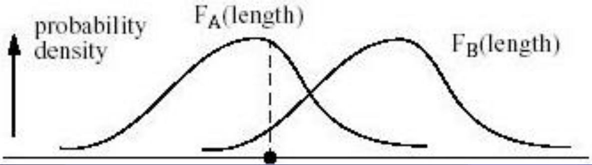
- Let $g \in \{\text{'f'}, \text{'m'}\}$ be a r.v. denoting the gender of a person.
- Let x be the measured length of the hair.

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 $P(g \mid x)$.

Solution:

Apply Bayes' Rule to get

$$\begin{aligned} P(g = \text{'m'} \mid x) &= \frac{P(x \mid g = \text{'m'})P(g = \text{'m'})}{P(x)} \\ &= \frac{P(x \mid g = \text{'m'})P(g = \text{'m'})}{P(x \mid g = \text{'f'})P(g = \text{'f'}) + P(x \mid g = \text{'m'})P(g = \text{'m'})} \end{aligned}$$

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

Selecting the most probably hypothesis

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

  Hypothesis with highest probability given observed data

$$
\begin{aligned}
t_{\text{MAP}} &= \arg\max_{t \in \mathcal{T}} P(t \mid \mathbf{x}) \\
&= \arg\max_{t \in \mathcal{T}} \frac{P(\mathbf{x} \mid t)\, P(t)}{P(\mathbf{x})} \\
&= \arg\max_{t \in \mathcal{T}} P(\mathbf{x} \mid t)\, P(t)
\end{aligned}
$$

Selecting the most probably hypothesis

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

  Hypothesis with highest probability given observed data

$$
\begin{aligned}
t_{\text{MAP}} &= \arg\max_{t\in\mathcal{T}} P(t \,|\, \mathbf{x}) \\
&= \arg\max_{t\in\mathcal{T}} \frac{P(\mathbf{x} \,|\, t)\, P(t)}{P(\mathbf{x})} \\
&= \arg\max_{t\in\mathcal{T}} P(\mathbf{x} \,|\, t)\, P(t)
\end{aligned}
$$

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

  Hypothesis with highest likelihood of generating observed data.

$$
t_{\text{MLE}} = \arg\max_{t\in\mathcal{T}} P(\mathbf{x} \,|\, t)
$$

  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.

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:**

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

Example: Cancer or Not?

Find MAP estimate:
When test returned a positive result,

$$t_{\mathsf{MAP}} = \arg \max_{t \in \{\text{disease, not disease}\}} P(t \mid +) = \arg \max_{t \in \{\text{disease, not disease}\}} P(+ \mid t)\, P(t)$$

Example: Cancer or Not?

Find MAP estimate:
When test returned a positive result,

$$t_{\mathsf{MAP}} = \arg \max_{t \in \{\text{disease, not disease}\}} P(t \,|\, +) = \arg \max_{t \in \{\text{disease, not disease}\}} P(+ \,|\, t)\, P(t)$$

Substituting in the correct values get

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

Therefore $y_{\mathsf{MAP}} =$ not disease.

Example: Cancer or Not?

Find MAP estimate:
When test returned a positive result,

$$t_{\text{MAP}} = \arg \max_{t \in \{\text{disease, not disease}\}} P(t \,|\, +) = \arg \max_{t \in \{\text{disease, not disease}\}} P(+ \,|\, t)\, P(t)$$

Substituting in the correct values get

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

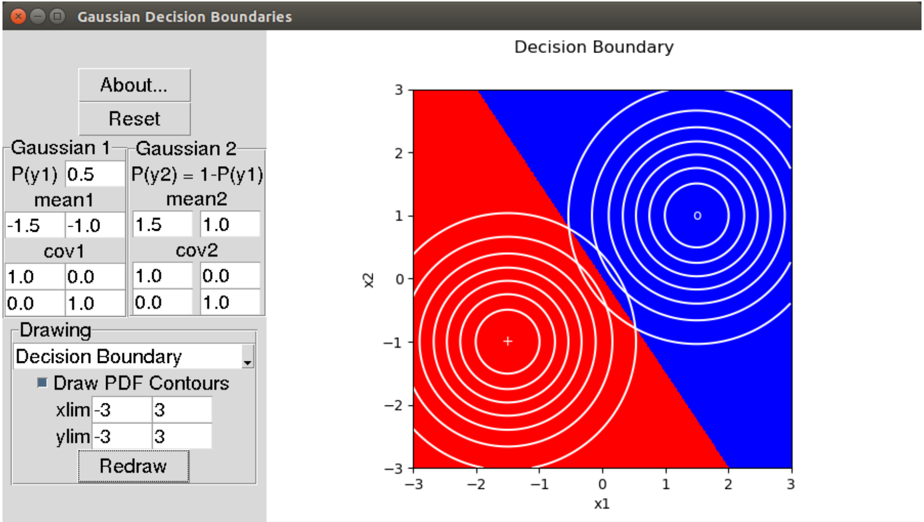
Therefore $y_{\text{MAP}} = \text{not disease}$.

The Posterior probabilities:

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

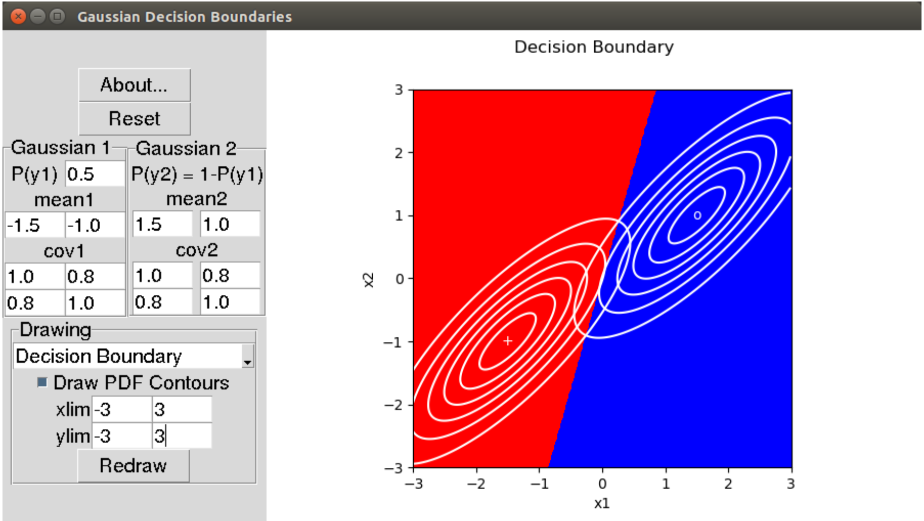
Demo: Gaussian Decision Boundaries

https://github.com/giampierosalvi/GaussianDecisionBoundaries



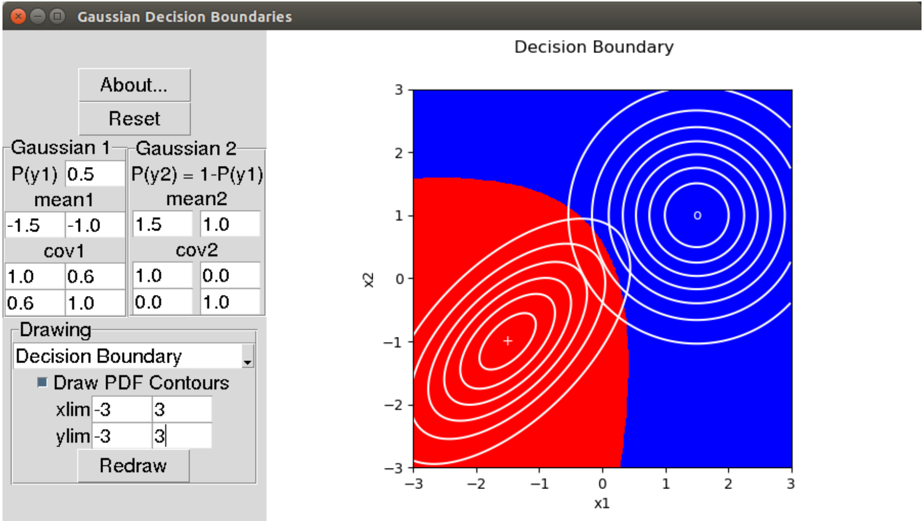
Demo: Gaussian Decision Boundaries

https://github.com/giampierosalvi/GaussianDecisionBoundaries



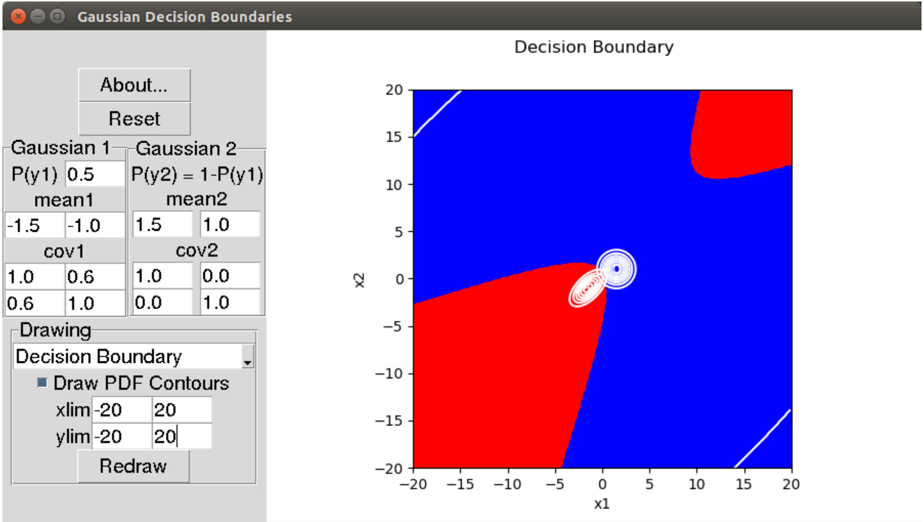
Demo: Gaussian Decision Boundaries

https://github.com/giampierosalvi/GaussianDecisionBoundaries



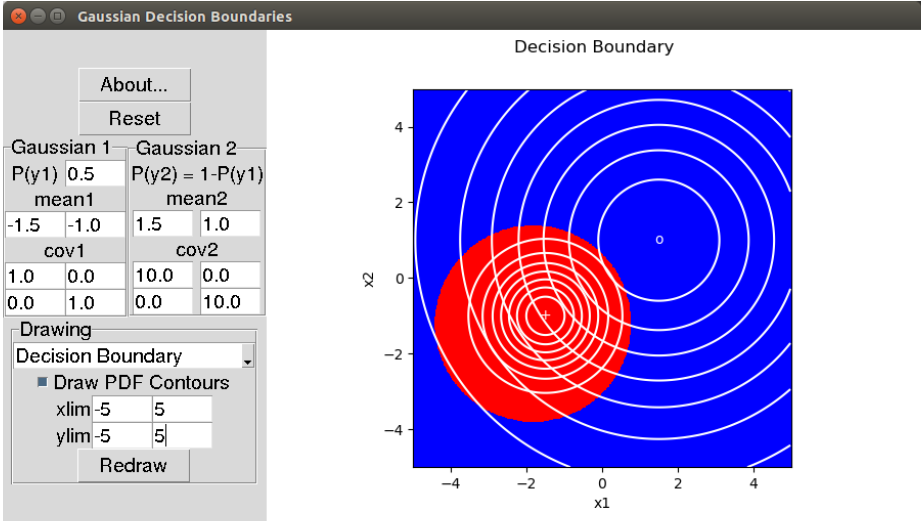
Demo: Gaussian Decision Boundaries

https://github.com/giampierosalvi/GaussianDecisionBoundaries



Demo: Gaussian Decision Boundaries

https://github.com/giampierosalvi/GaussianDecisionBoundaries



Inference vs Decision

- Inference: estimate posterior $p(c_k|\mathbf{x})$
- Decision: use $p(c_k|\mathbf{x})$ to make class assignments
- we could go directly from x to c_k with discriminant functions

Generative vs Discriminative

Generative:

- estimate $p(x|c_k)$ and $p(c_k)$
- compute $p(c_k|\mathbf{x}) = \frac{p(x|c_k)p(c_k)}{\sum_j p(x|c_j)p(c_j)}$
- use decision theory

Discriminative:

- estimate posterior $p(c_k|x)$ directly
- use decision theory
- we could also use discriminant functions $f_k(x)$ with no probabilistic interpretation

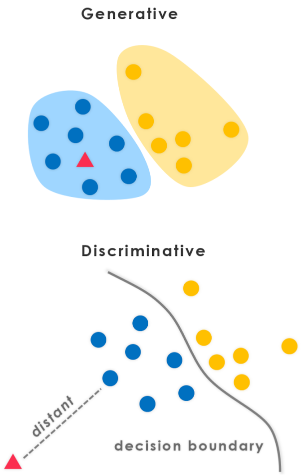


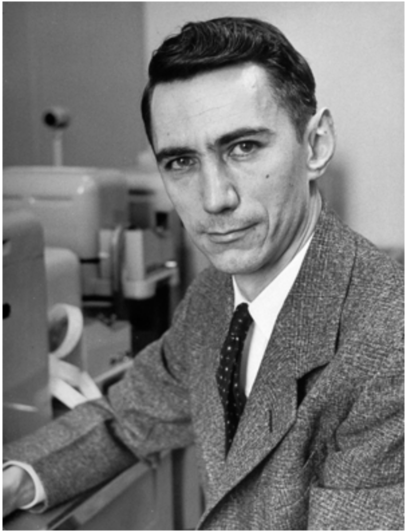
Figure from Nguyen *et al.* 2015. `http://www.evolvingai.org/fooling`

Outline

1. Probability Theory Reminder
   - Axioms and Properties
   - Common Distributions
   - Moments
2. Probabilistic Machine Learning
   - Supervised Learning, General Definition
   - Regression
   - Classification
   - Bayes decision theory
3. Information Theory

Information Theory

- Pioneered by Shannon in the 1940s
- probabilities describe uncertainty of events
- we are interested in the information gained observing the outcome of an event



Information

$$h(x) = -\log_2 p(x)$$

Properties of Information

$$h(x) = -\log_2 p(x)$$

1. If x and y are independent variables:

$$\begin{aligned} p(x,y) &= p(x)p(y) \\ h(x,y) &= h(x) + h(y) \end{aligned}$$

2. The sure event Ω associated with zero information:

$$p(\Omega) = 1 \quad \Rightarrow \quad h(\Omega) = 0$$

3. If $p(x) < p(y) \quad \Rightarrow \quad h(x) > h(y)$

Information and transmissions

Suppose we want to send the outcome of $x \in \{x_1, x_2, \dots\}$ (finite or countable) with distribution $p(x)$

Entropy

The expected number of bits of information is

$$\mathbb{E}[h(x)] = \sum_i p(x_i)h(x_i) = -\sum_i p(x_i)\log_2 p(x_i) = H(x)$$

Example

Suppose we want to send the outcome of $x \in \{a, b, c, d, e, f, g, h\}$

Uniform distribution: $p(x_i) = \frac{1}{8}$

Then $H(x) = -\sum_{i=1}^{8} \frac{1}{8} \log_2 \frac{1}{8} = 3$ bits

Possible code:

$$
\begin{aligned}
a &= 000 \\
b &= 001 \\
&\dots \\
h &= 111
\end{aligned}
$$

Example

Suppose we want to send the outcome of $x \in \{a, b, c, d, e, f, g, h\}$

Non uniform distribution: $p(x) = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\}$

Then $H(x) = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 2^2 + \frac{1}{8}\log_2 2^3 + \frac{1}{16}\log_2 2^4 + 4\frac{1}{64}\log_2 2^6 = 2$ bits

Possible code:

$$
\begin{aligned}
a &= 0 & e &= 111100 \\
b &= 10 & f &= 111101 \\
c &= 110 & g &= 111110 \\
d &= 1110 & h &= 111111
\end{aligned}
$$

Noiseless coding theorem (Shannon 1948)

Theorem

The entropy is the lower bound to the number of bits required on average to transmit the state of a random variable.

Other interpretations:

- Thermodynamics: measure of equilibrium
- Statistical mechanics: measure of disorder

Units:

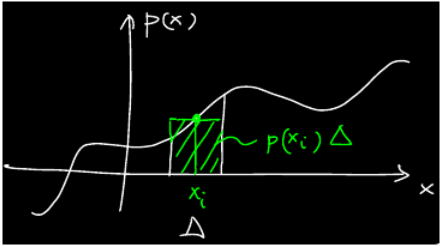
- $\log_2 \rightarrow$ bits
- $\ln \rightarrow$ nats (from natural log)

Entropy and continuous variables

If $x \in \mathbb{R}$ and $p(x) \in C^0$ is a continuous function in $\mathbb{R}$

$$
\begin{aligned}
H_\Delta &= -\sum_i p(x_i)\Delta \ln\left(p(x_i)\Delta\right) \\
&= -\sum_i p(x_i)\Delta \ln p(x_i) - \ln\Delta \xrightarrow[\Delta\to 0]{} \infty
\end{aligned}
$$

Mean value theorem



$$\int_{i\Delta}^{(i+1)\Delta} p(x)dx = p(x_i)\Delta$$

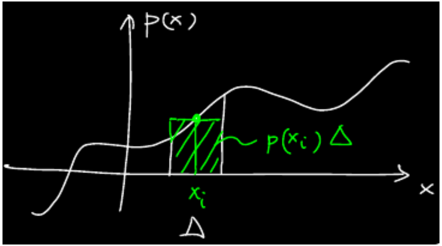
Entropy and continuous variables

If $x \in \mathbb{R}$ and $p(x) \in C^0$ is a continuous function in $\mathbb{R}$

$$\begin{aligned} H_\Delta &= -\sum_i p(x_i)\Delta \ln\left(p(x_i)\Delta\right) \\ &= -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta \xrightarrow[\Delta\to 0]{} \infty \end{aligned}$$

Infinite information to transmit a real number!

Mean value theorem



$$\int_{i\Delta}^{(i+1)\Delta} p(x)dx = p(x_i)\Delta$$

Differential Entropy

$$H_\Delta = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta \xrightarrow[\Delta\to 0]{} \infty$$

- remove the offending term $\ln \Delta$
- take the limit for $\Delta \to 0$

Differential Entropy

$$H(x) = \lim_{\Delta\to 0} -\sum_i p(x_i)\Delta \ln p(x_i) = -\int p(x) \ln p(x) dx$$

Maximum Entropy

What are the distributions that give maximum (differential) entropy?

- Discrete case (k values): uniform distribution,

$$H = \ln k$$

- Continuous case: Gaussian distribution,

$$H = \frac{1}{2}\left[1 + \ln(2\pi\sigma^2)\right]$$

Conditional Entropy

Given two variables x and y:

$$\begin{aligned} H(y|x) &= -\int\int p(x,y)\ln p(y|x)dxdy \\ &= H(x,y) - H(x) \end{aligned}$$

Relative Entropy (Kullback-Leibler divergence)

Given two distributions $p(x)$ and $q(x)$
Cross entropy:

$$\mathbb{E}_{x\sim p}[-\ln q] \quad = \quad -\int p(x)\ln q(x)dx$$

Kullback-Leibler divergence:

$$\begin{aligned}
\mathsf{KL}(p||q) &= \mathbb{E}_{x\sim p}[-\ln q] - \mathbb{E}_{x\sim q}[-\ln p] \\
&= -\int\int p(x)\ln q(x)dx + \int\int q(x)\ln p(x)dx \\
&= -\int\int p(x)\ln\frac{q(x)}{p(x)}dx
\end{aligned}$$

Mutual Information

Given two variables x and y, are they independent?
Mutual information:

$$
\begin{aligned}
I(x, y) &= \mathsf{KL}(p(x, y) || p(x)p(y)) \\
&= -\int\int p(x, y) \ln \frac{p(x)p(y)}{p(x, y)} dx dy
\end{aligned}
$$

- $I(x, y) \geq 0$
- $I(x, y) = 0 \Leftrightarrow x$ and y are independent

$$
I(x, y) = H(x) - H(x|y) = H(y) - H(y|x)
$$