Lecture 5: Probabilistic Reasoning

DD2421

Bob L. T. Sturm

My three lecture block

- Today: probabilistic reasoning
- Next time: learning as inference
- Next next time: learning with latent variables

My three lecture block

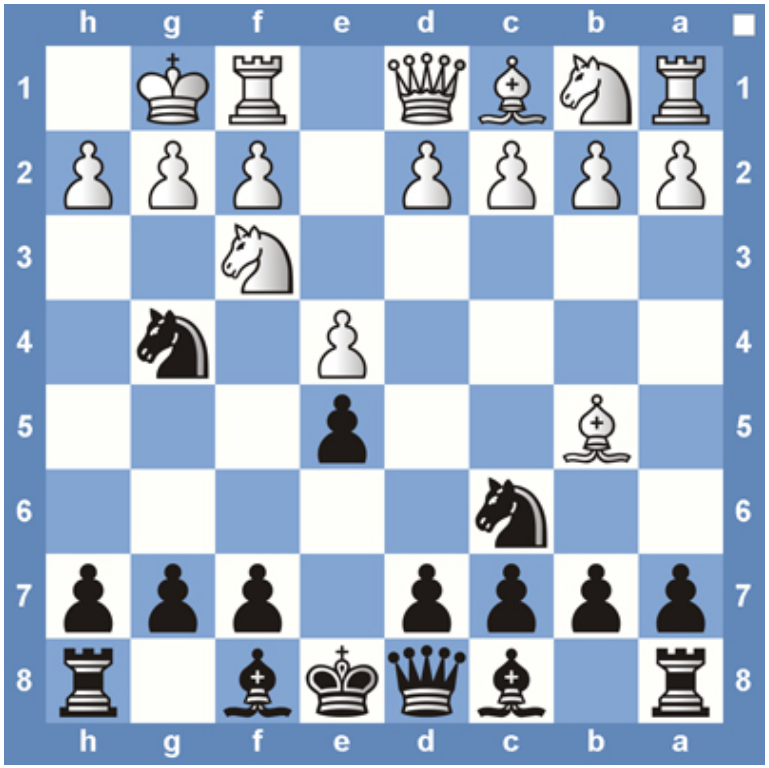
- Today: probabilistic reasoning
- Next time: learning as inference
- Next next time: learning with latent variables

Constructive alignment

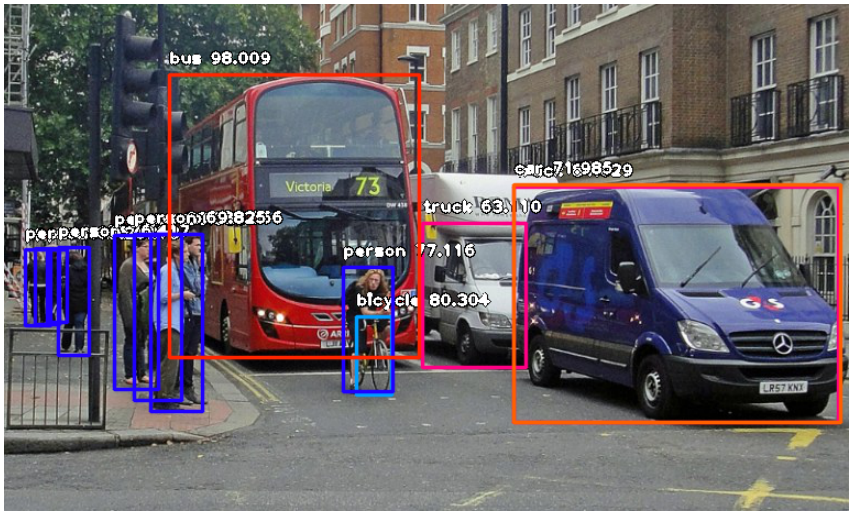
- assignments: part of lab 3 is relevant (Naïve Bayes)
- two exam problems are on probabilistic methods
- the topic is fundamental for the advanced course (DD2434)

Why Machine Learning?

AI in the 1970s



AI today



We need to deal with uncertainty!

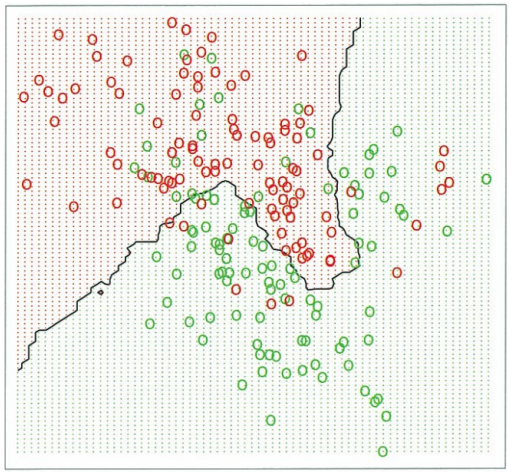
Uncertainty

- How can we deal with the randomness we see around us?
- How can we formally describe degrees of belief?

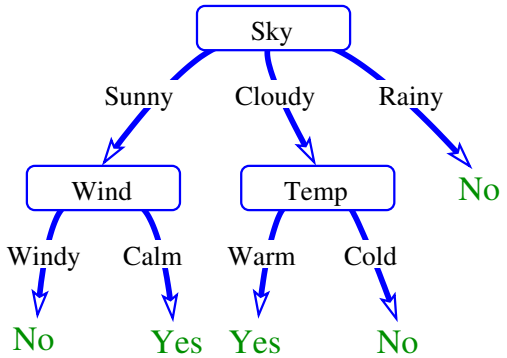
Example: Monthy Hall Problem

Examples of ML Methods seen so far

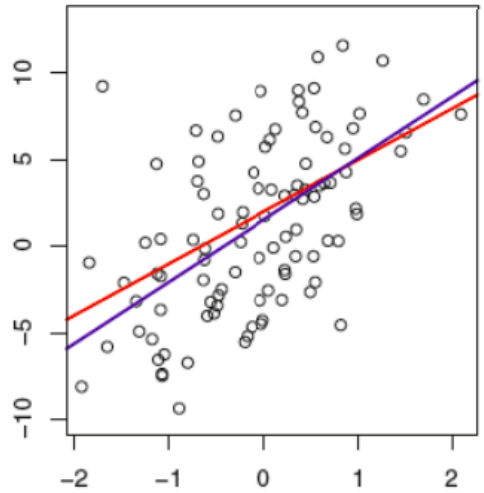
k Nearest Neighbour



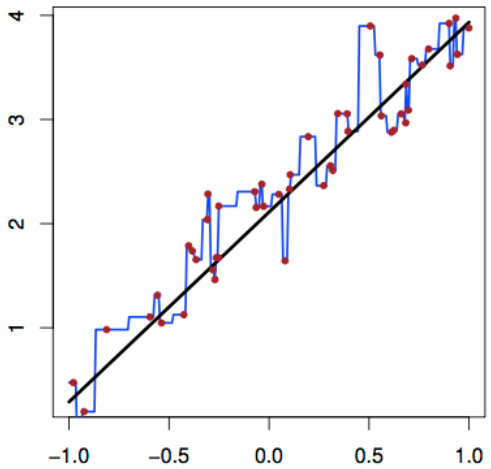
Decision Trees



Least squares Regression



k-NN Regression



Probability Theory in ML

Incorporate probabilistic thinking at all levels

- We start with incomplete knowledge (uncertainty)
- We then reduce uncertainty by incorporating observations
- We then propagate belief and update our estimations.

Probabilistic machine learning provides
something like a unified theory for ML!

Advantages of Probability Based Methods

- **Interpretability!** Can be more transparent and mathematically rigorous than other ML methods. (But humans can have problems interpreting probability, as we will see later today.)
- **Transparency!** Assumptions can be made more explicit than in other methods (we will see this when we look at probabilistic linear regression next time).
- **Efficiency!** Can work in regimes that are data poor.
- **Flexibility!** Easy to merge different parts of a complex system and to update current knowledge with new observations.
- **Encompassing!** Aspects of learning and inference can be cast under the same theory.

Disadvantages of Probability Based Methods

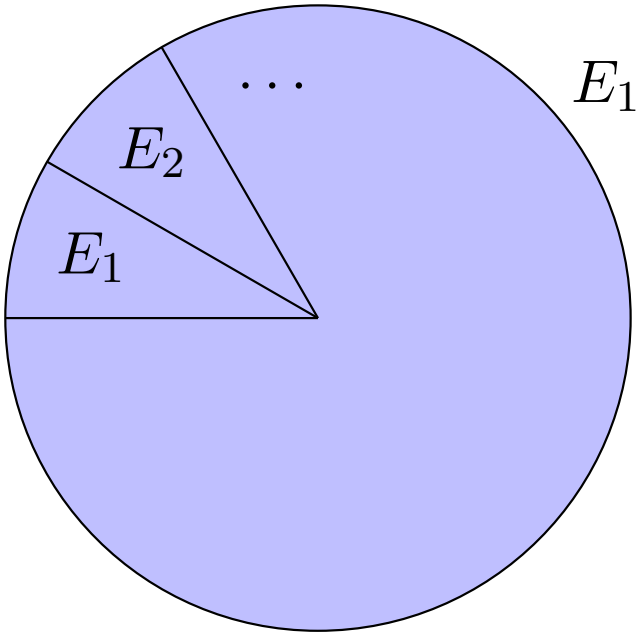
- **Complicated ...** Often hard to derive closed solutions. Need to resort to computation and heuristic approximations.
- **Scalability ...** Not computationally scalable to large datasets (but many argue that the need for large data sets is a disadvantage).

Outline

Axiomatic definition of probabilities (Kolmogorov)

Given a *sample space* Ω of all possible *outcomes*, and an *event* E, or set of outcomes from Ω (possibly empty), then:

1. $P(E) \geq 0$ for all $E \subseteq \Omega$
2. If $E = \Omega$ then $P(\Omega) = 1$ (sure event) $\Rightarrow 0 \leq P(E) \leq 1$
3. If $E_1, E_2, \ldots$ is a countable sequence of pairwise disjoint events, then

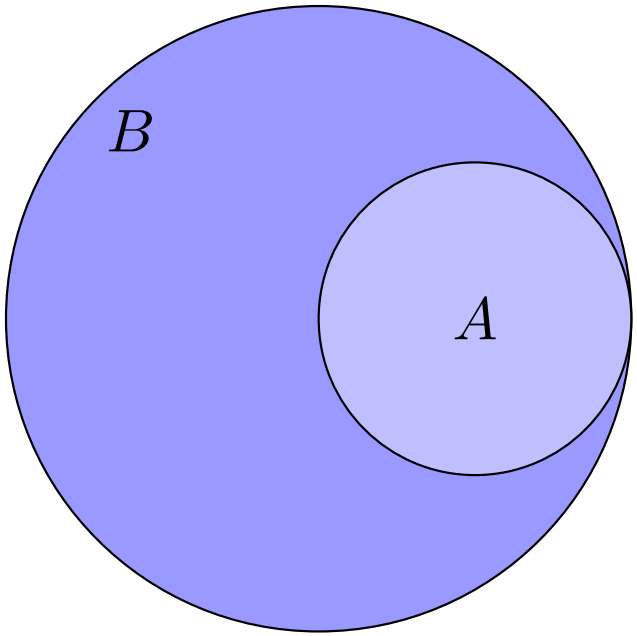


$$E_1 \cup E_2 \cup \cdots$$

$$P(E_1 \cup E_2 \cup \cdots) = \sum_i P(E_i)$$

Consequences

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

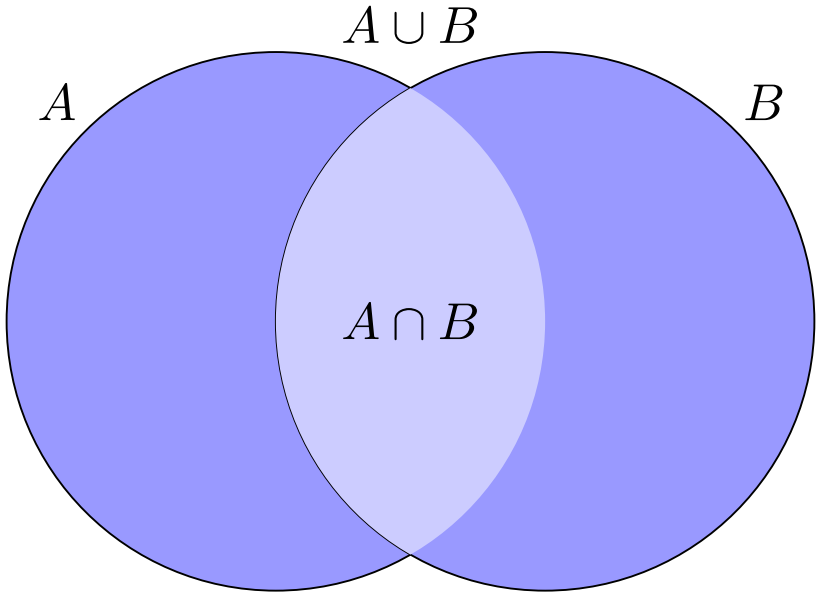


Example (die): $A = \{3\}, \quad B = \{\text{odd}\}$

2. Empty set $\emptyset$: $P(\emptyset) = 0$
Example (die): $P(A \cap B)$ where $A = \{\text{odd}\}, B = \{\text{even}\}$

More Consequences: Addition

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

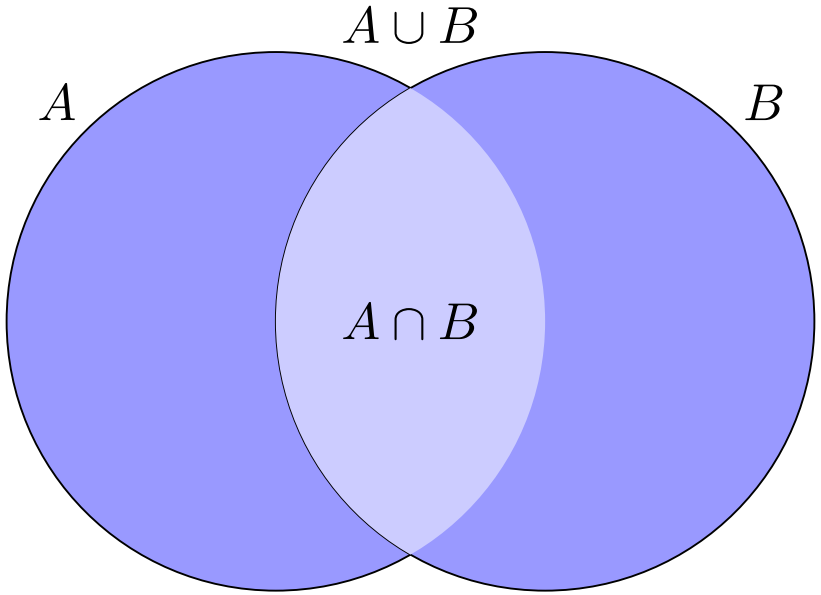


Example:

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

More Consequences: Addition

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

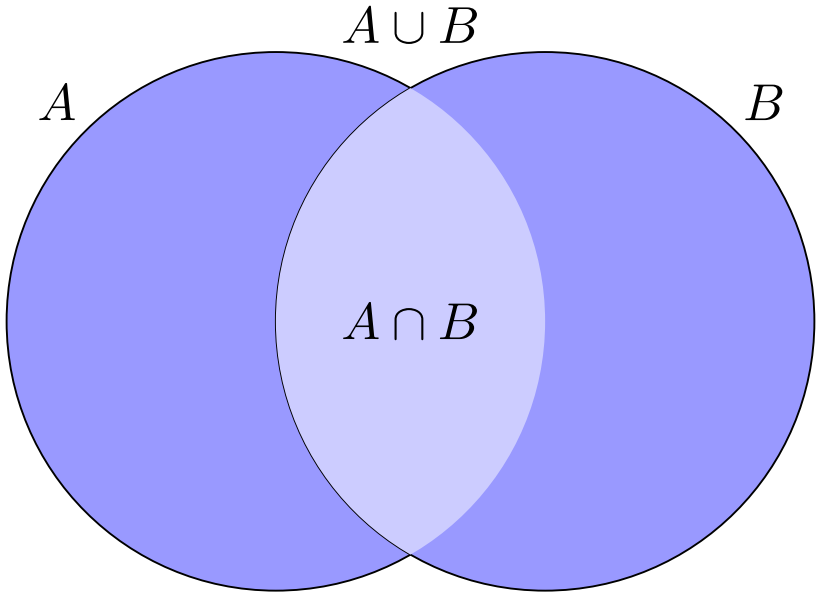


Example:

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

More Consequences: Addition

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

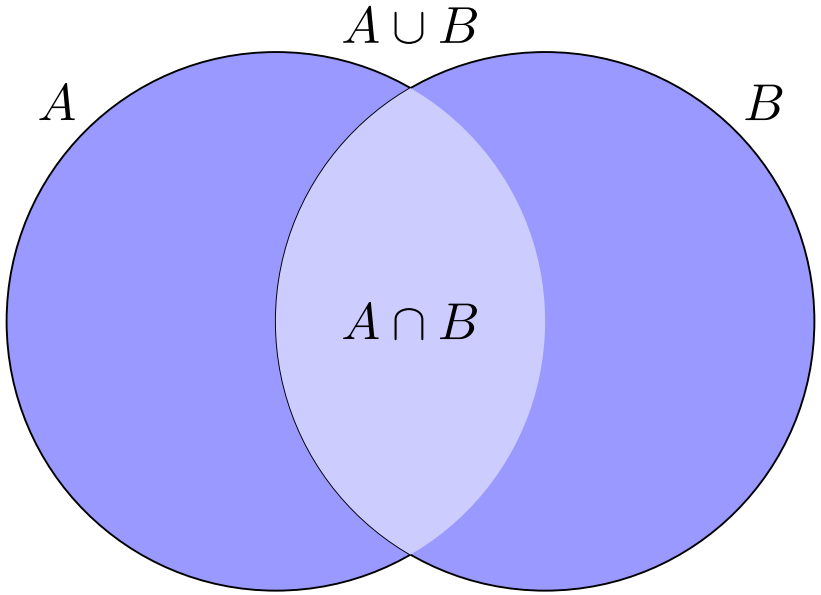


Example:

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

More Consequences: Addition

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

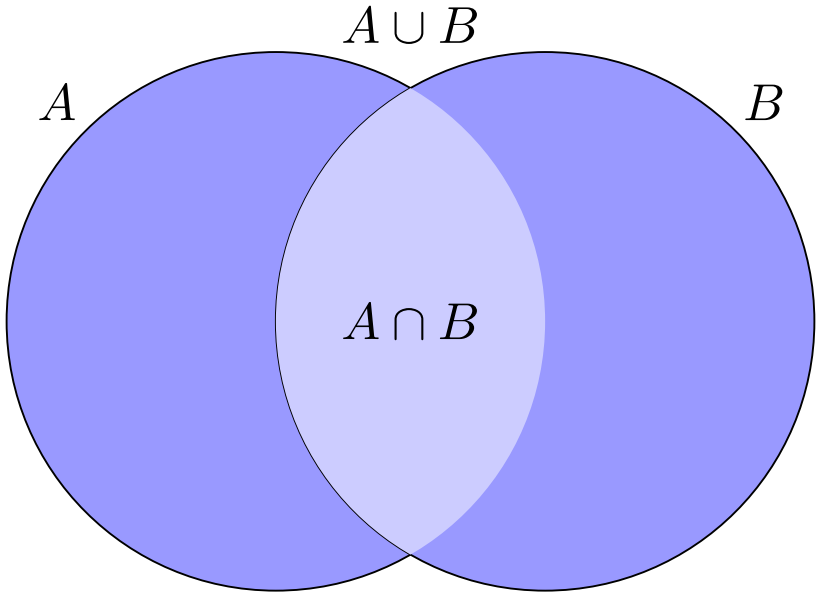


Example:

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

More Consequences: Addition

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

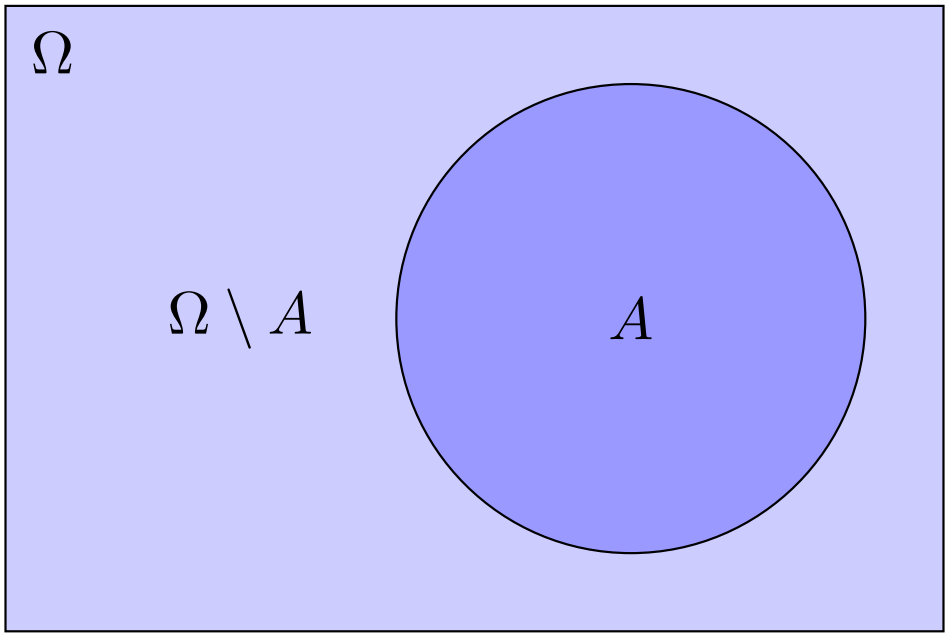


Example:

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

More Consequences: Negation

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

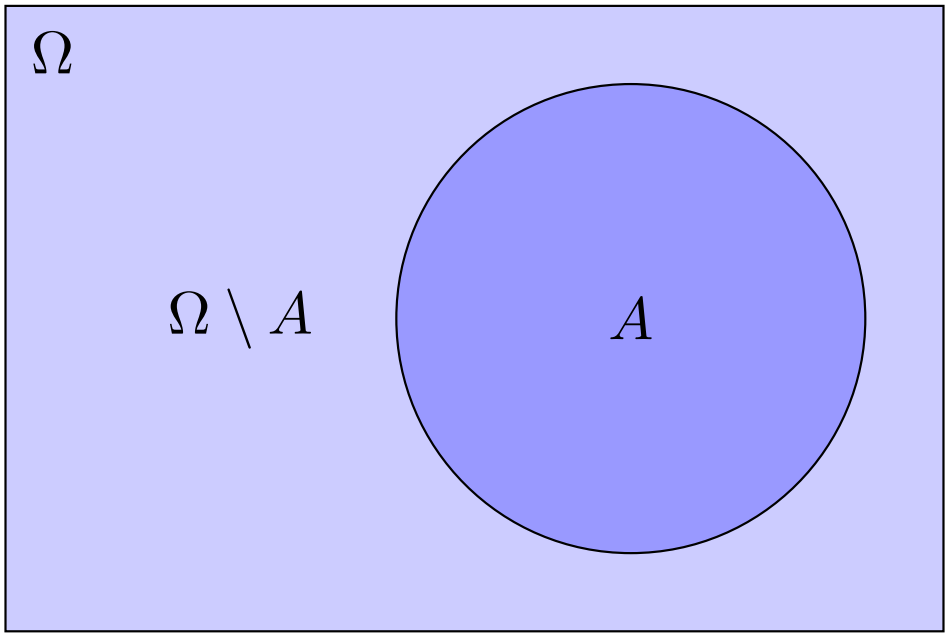


Example:

$$A = \{1, 2\}, \qquad P(A) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$
$$\bar{A} =$$

More Consequences: Negation

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

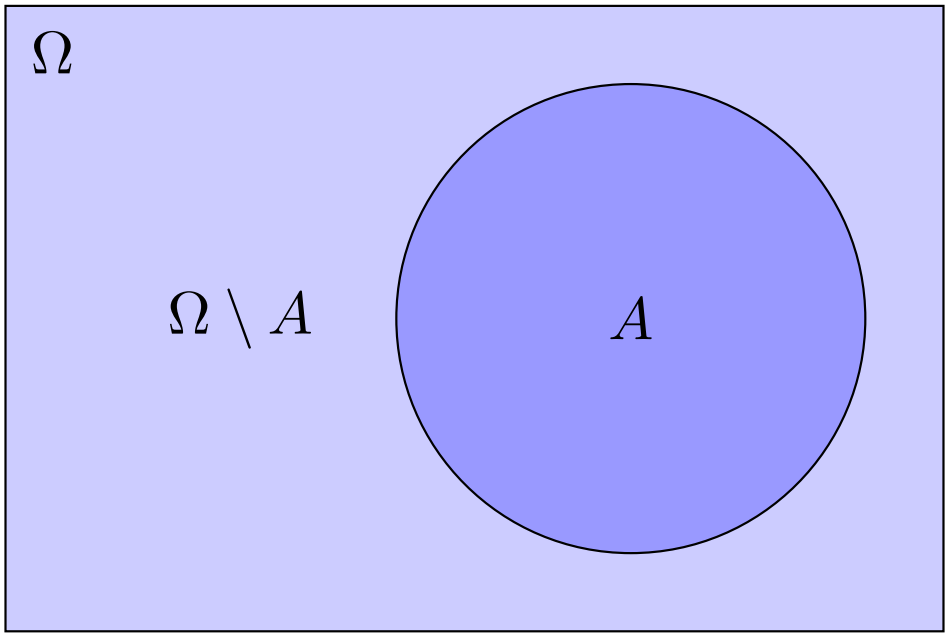


Example:

$$\begin{array}{rclrcl} A &=& \{1,2\}, & P(A) &=& \frac{1}{6}+\frac{1}{6}=\frac{1}{3} \\ \bar{A} &=& \{3,4,5,6\}, & P(\bar{A}) &=& \end{array}$$

More Consequences: Negation

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



Example:

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

Random (Stochastic) Variables

- Sample spaces (set of all possible outcomes) are not numbers, e.g., die rolls, heads/tails, playing cards, words, etc.
- We need to "convert" outcomes to numbers.
- Let's map the set of outcomes to a set of numbers!

Random (Stochastic) Variables

- Sample spaces (set of all possible outcomes) are not numbers, e.g., die rolls, heads/tails, playing cards, words, etc.
- We need to "convert" outcomes to numbers.
- Let's map the set of outcomes to a set of numbers!

A *random variable (r.v.)* is neither random nor a variable. It's a function that does exactly what we need, e.g.,

$$X : \Omega \to \mathbb{R}$$

Probability distributions

The *probability distribution function (PDF)* of a r.v. maps its range to positive numbers

$$Pr(x) : X \rightarrow \mathbb{R}_{>0}$$

- $Pr(x)$ describes how probability *density* is distributed over the range of X, e.g., the probability of $0 \leq X \leq 1$ is given by:

$$P[0 \leq X \leq 1] = \int_{x=0}^{1} Pr(x)dx$$

- "X is distributed $Pr(x)$" written more compactly:

$$X \sim Pr(x)$$

Examples

Student in HT19 machine learning course hands in exam, which is then graded.

Ω	–	All possible graded exams of all students in class
$X : \Omega \to \mathbb{R}$	–	Random variable mapping exam outcome to score
$X \sim Pr(x)$	–	Gaussian distribution, $\mu = 32.5$, $\sigma^2 = 75.5$

Examples

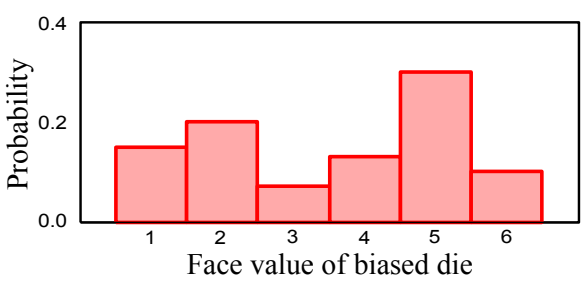
Student in HT19 machine learning course hands in exam, which is then graded.

Ω	–	All possible graded exams of all students in class
$X : \Omega \to \mathbb{R}$	–	Random variable mapping exam outcome to score
$X \sim Pr(x)$	–	Gaussian distribution, $\mu = 32.5$, $\sigma^2 = 75.5$

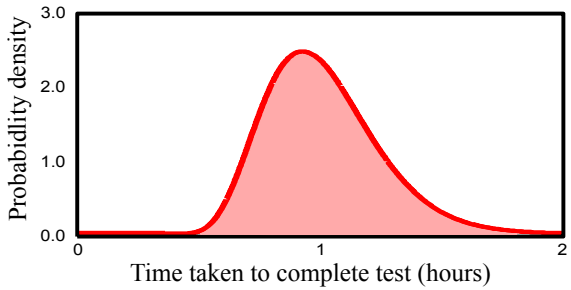
Snälla: Work hard so that μ is much higher!

Types of Random Variables

- The range of a **discrete random variable** is a countable set, e.g., $\{1, 2, \ldots, 6\}$.
- The range of a **continuous random variable** is an uncountable set, e.g., the reals $\mathbb{R}$, or $[0, 1)$.



discrete pdf



continuous pdf

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

Examples of Random Variables

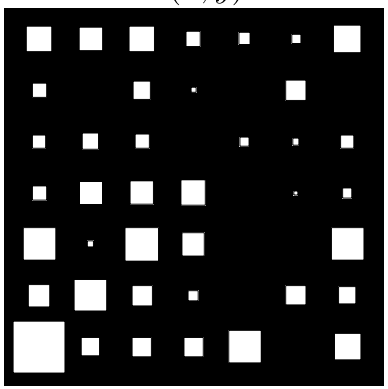
- Discrete events: either 1, 2, 3, 4, 5, or 6.
- Probability *mass* function (PMF) $P[X = d] = Pr(d)$
- $P[X = d] = 1/6, d \in \{1, 2, 3, 4, 5, 6\}$ (fair die)

- Any real number (theoretically infinite)
- PDF $Pr(x)$
- $P[t_1 < X < t_2]$ involves integrating $Pr(x)$
- $P[X = t_1] = 0$ though $Pr(t_1) \neq 0$

Joint Probabilities

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

$Pr(x, y)$



← **discrete joint pdf**

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

Marginalization

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

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

and integrating for the continuous case

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

Marginalization (cont.)

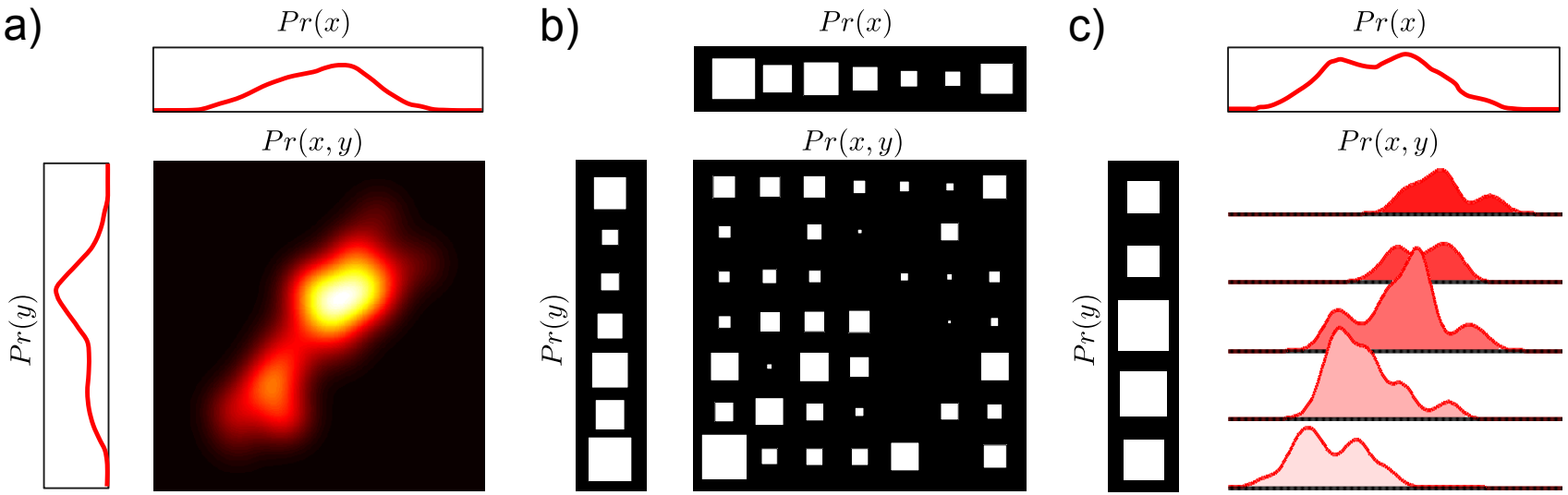


Figure from **Computer Vision: models, learning and inference** by Simon Prince.

Conditional Probabilities

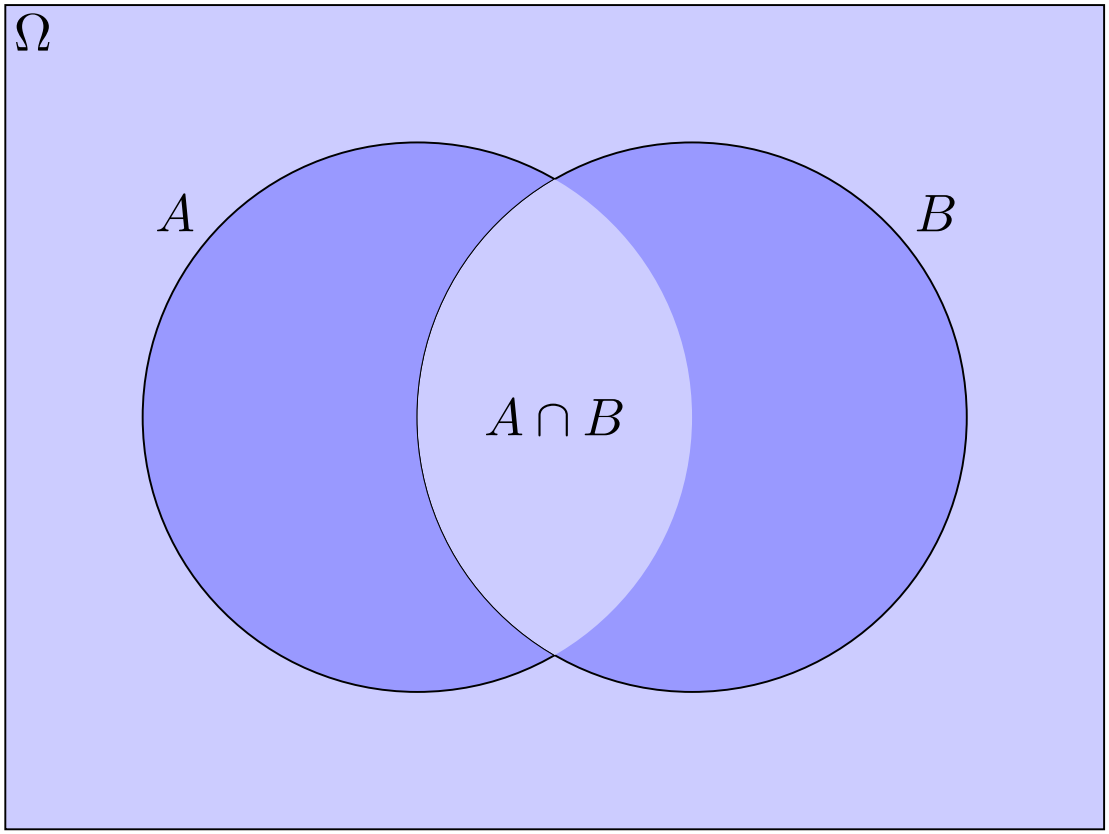
$$P(A|B)$$

The probability of event A *given* that event B has happened

Note: This is not the probability that event A *and* B will happen

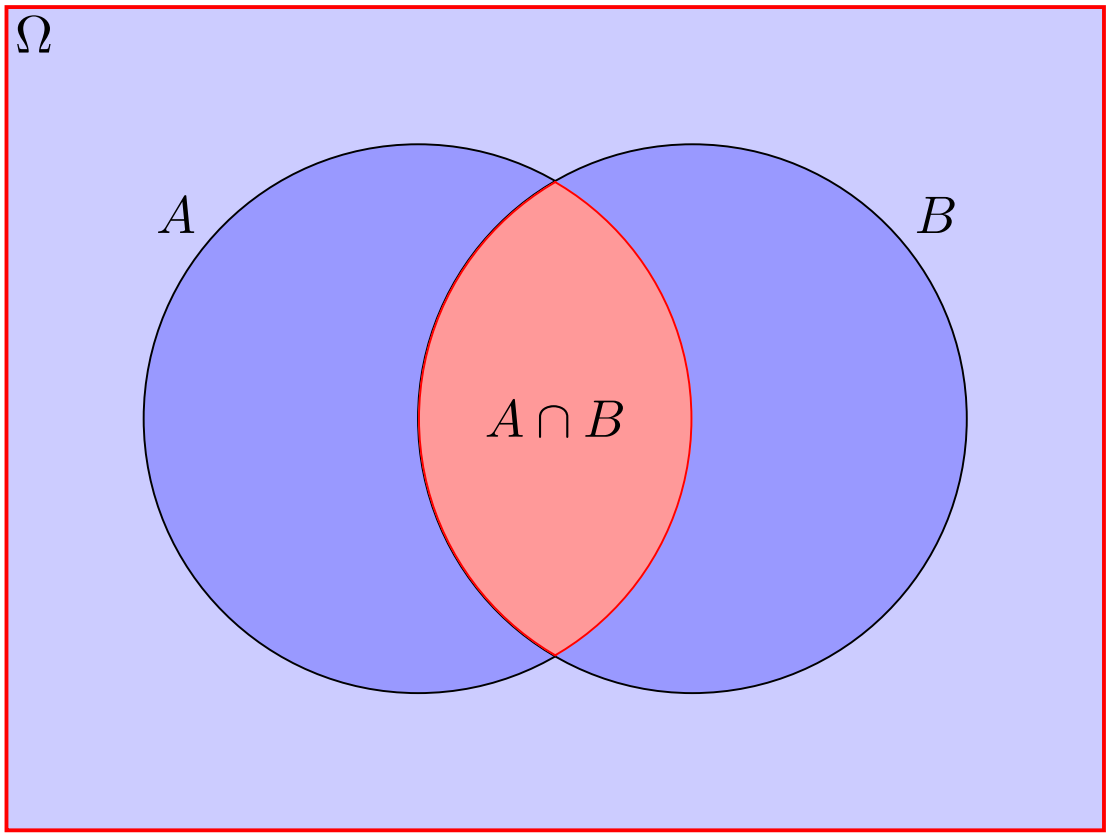
Conditional Probabilities

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



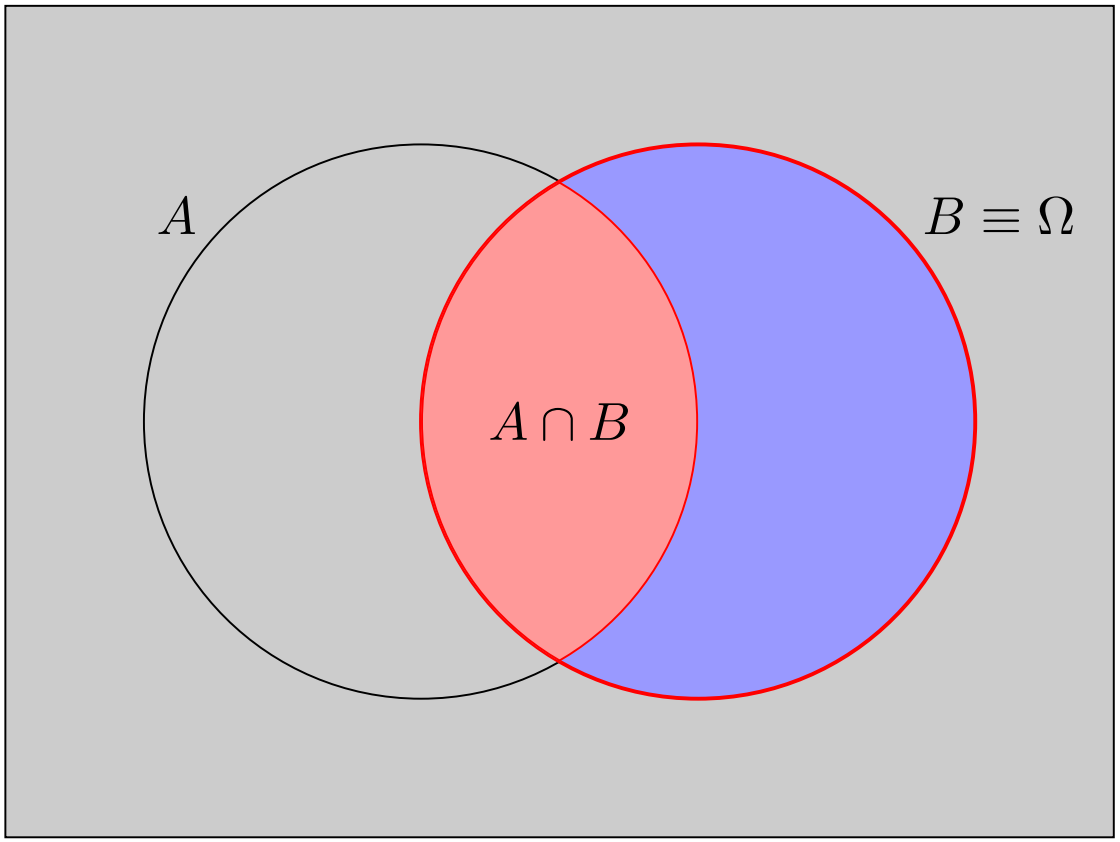
Conditional Probabilities

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



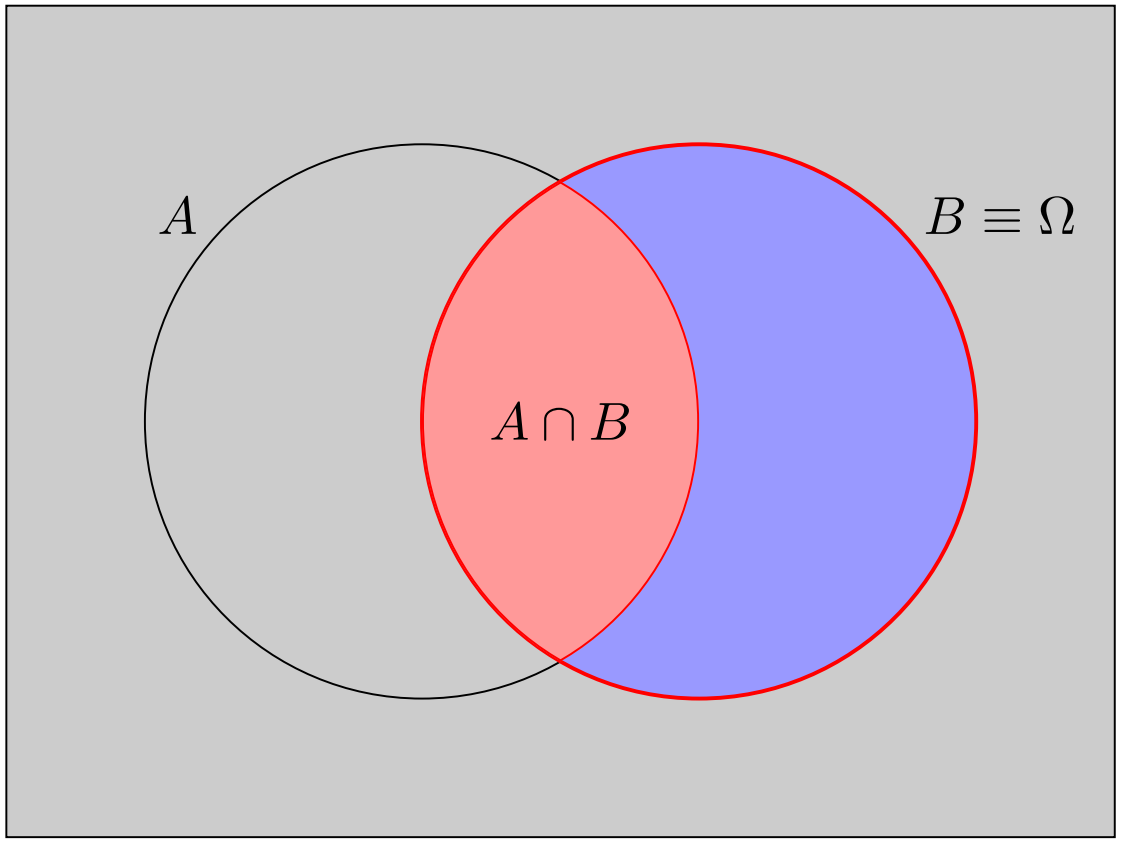
Conditional Probabilities

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



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 density for different values of r.v. X given that Y is fixed to value y.
- The conditional probability can be found from the joint distribution function $Pr(x, y)$:

$$Pr(x \,|\, Y = y) = \frac{Pr(x, Y = y)}{Pr(Y = y)} = \frac{Pr(x, Y = y)}{\int_x Pr(x, Y = y)\, dx}$$

- Extract an appropriate slice, and then normalize it.

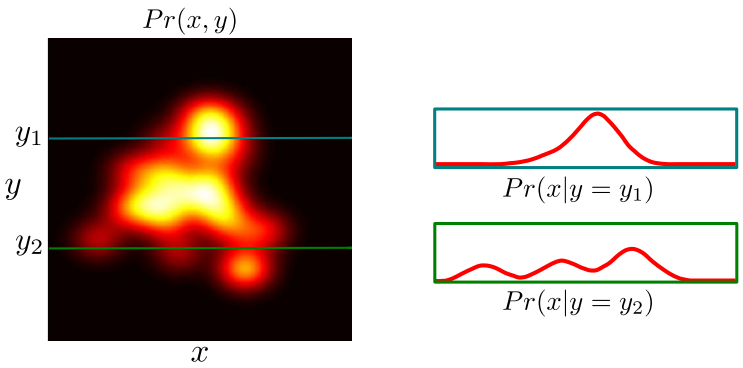


Figure from **Computer Vision: models, learning and inference** by Simon Prince.

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:

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

B happening tells us nothing about A happening

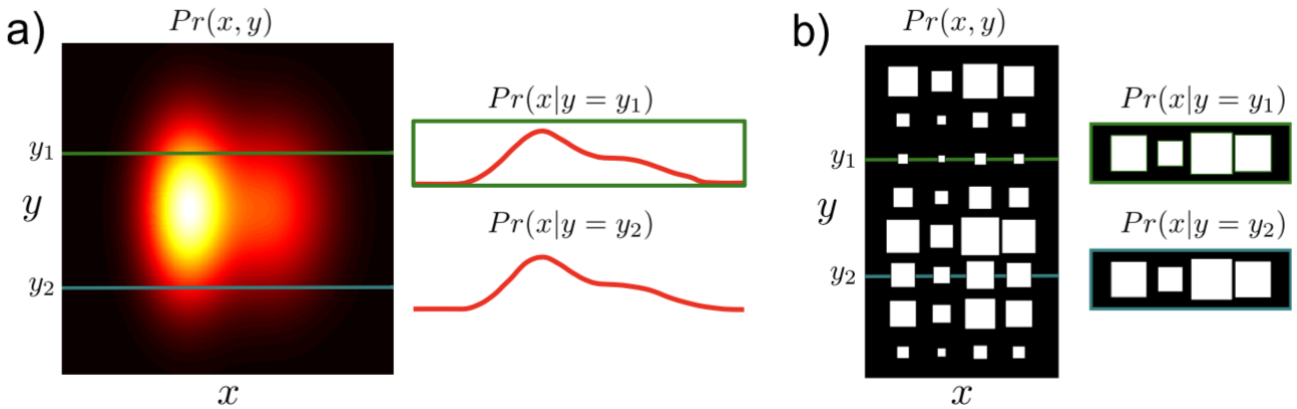


Figure from **Computer Vision: models, learning and inference** by Simon Prince.

Bayes' Theorem

if

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

then

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

Bayes' Theorem

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' Theorem

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

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

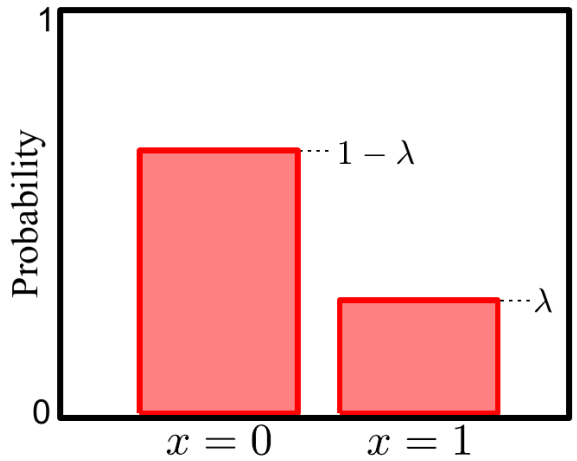
Memorize this and its derivation!

Bernoulli: binary variables

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

Then $P[x = 0] = 1 - \lambda$, and

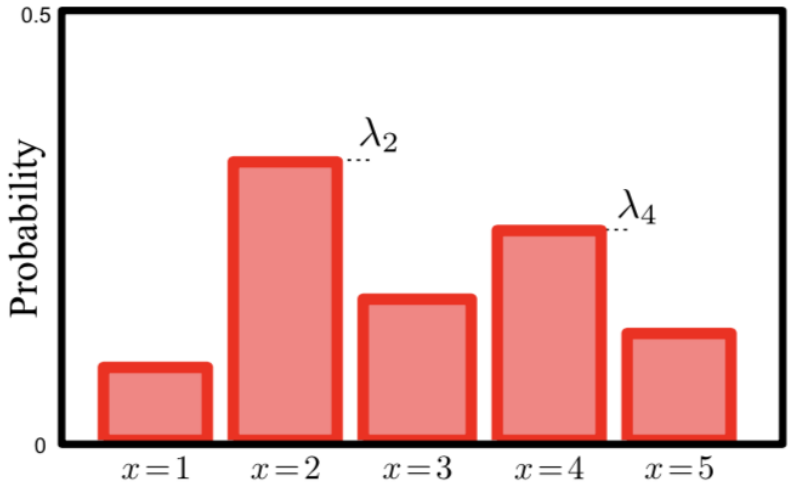
$$Pr(x) = \lambda^x (1 - \lambda)^{1-x} = \begin{cases} \lambda, & \text{if } x = 1, \\ 1 - \lambda, & \text{if } x = 0 \end{cases}$$



Categorical

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

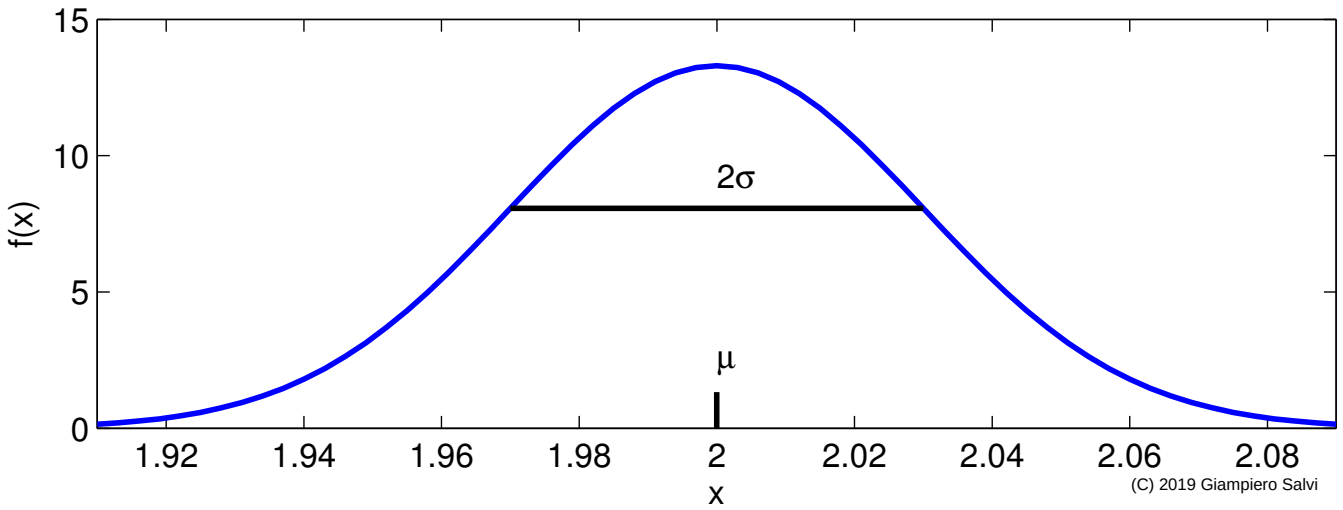
$$P[x_k = 1] = \lambda_k$$



Gaussian distributions: One-dimensional

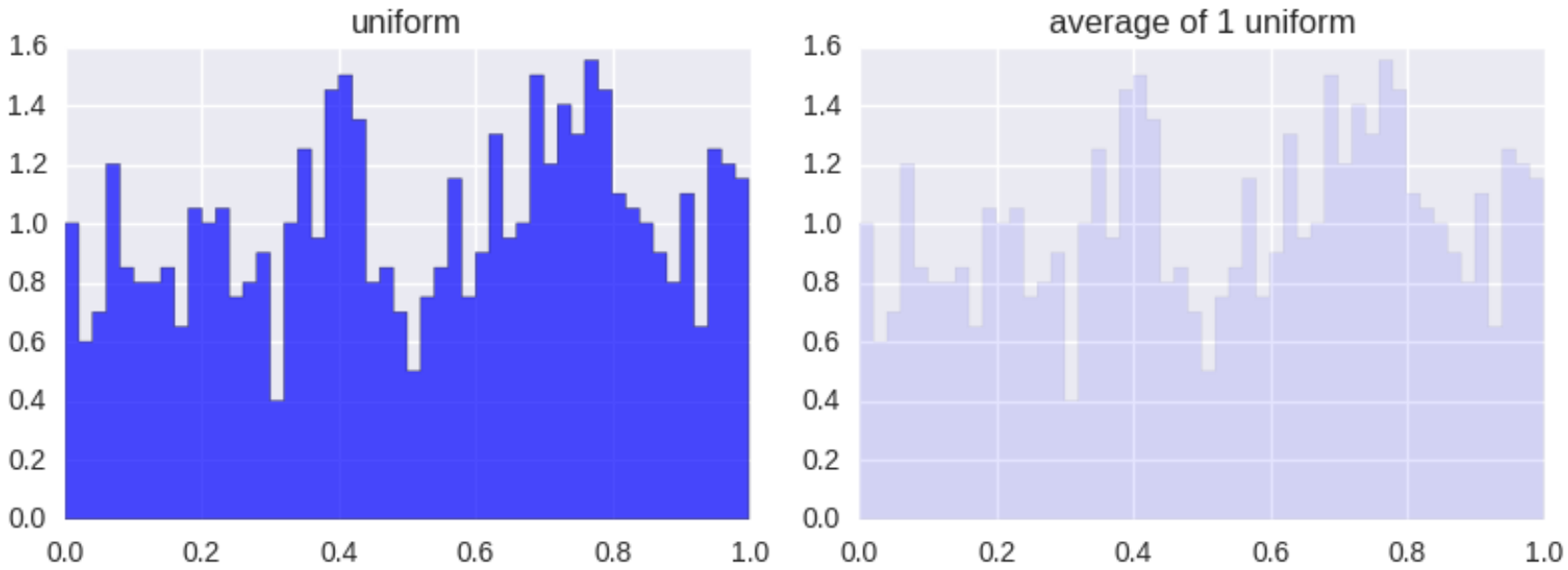
- Univariate normal distribution
- Domain: real numbers ($x \in \mathbb{R}$)

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



Why Gaussian: Central Limit Theorem

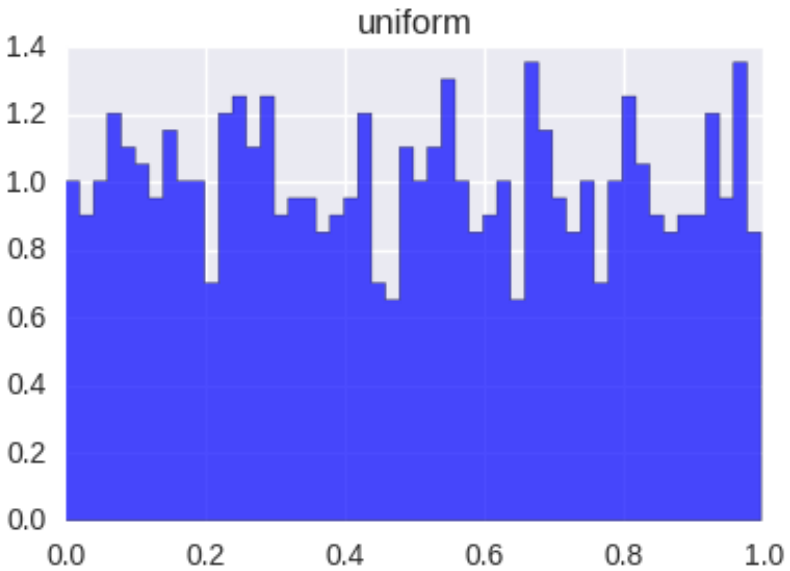
The distribution of a linear combination of a large number of independent, identically distributed (iid) variables will tend to normal, regardless of the underlying distribution.[1]

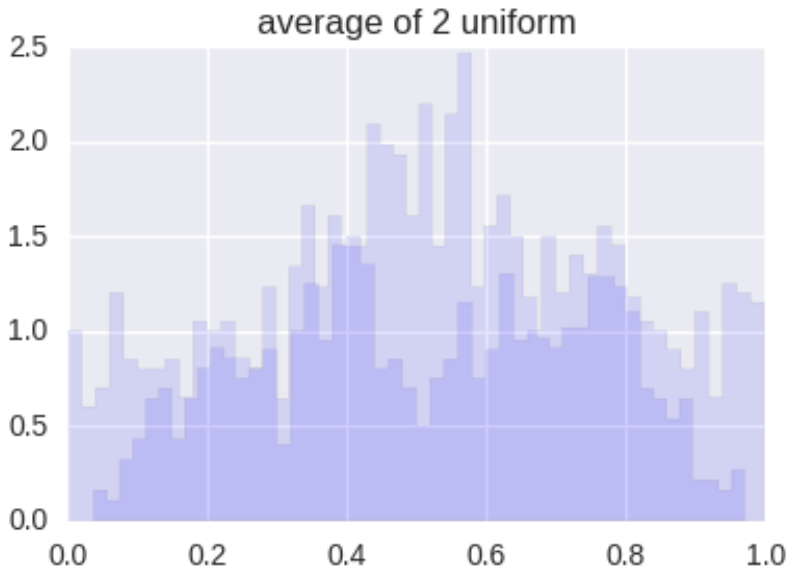


[1] **Bishop:2006ui**.

Why Gaussian: Central Limit Theorem

The distribution of a linear combination of a large number of independent, identically distributed (iid) variables will tend to normal, regardless of the underlying distribution.[1]

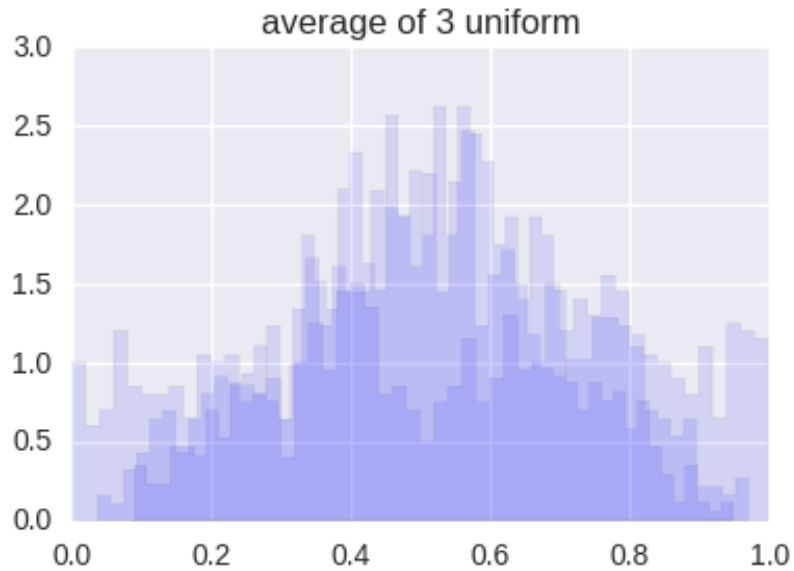




[1] **Bishop:2006ui**.

Why Gaussian: Central Limit Theorem

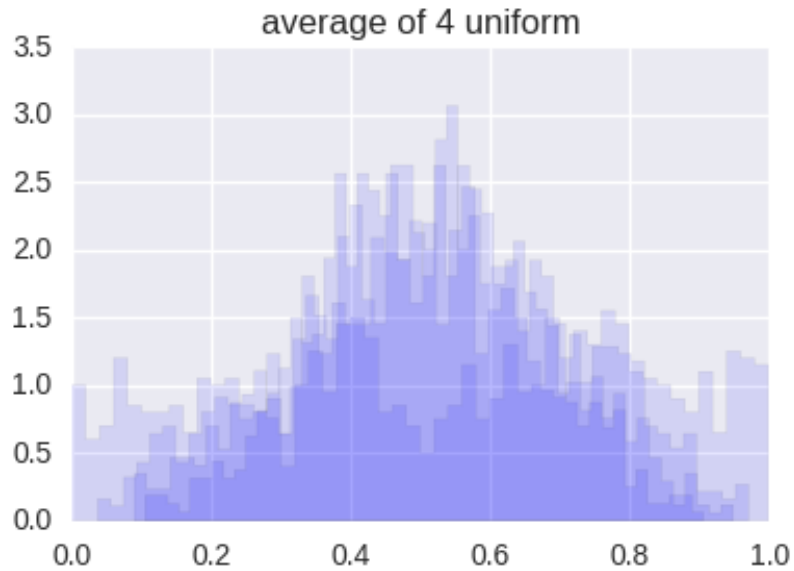
The distribution of a linear combination of a large number of independent, identically distributed (iid) variables will tend to normal, regardless of the underlying distribution.[1]



[1] **Bishop:2006ui**.

Why Gaussian: Central Limit Theorem

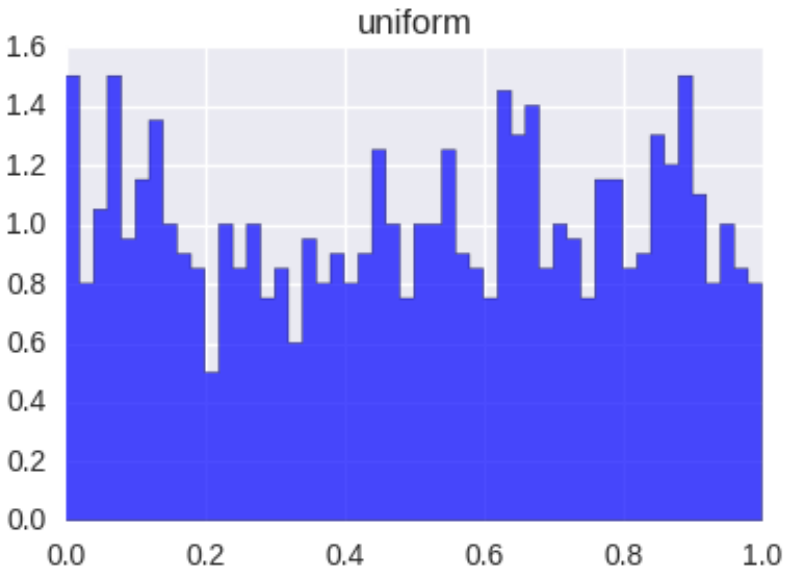
The distribution of a linear combination of a large number of independent, identically distributed (iid) variables will tend to normal, regardless of the underlying distribution.[1]

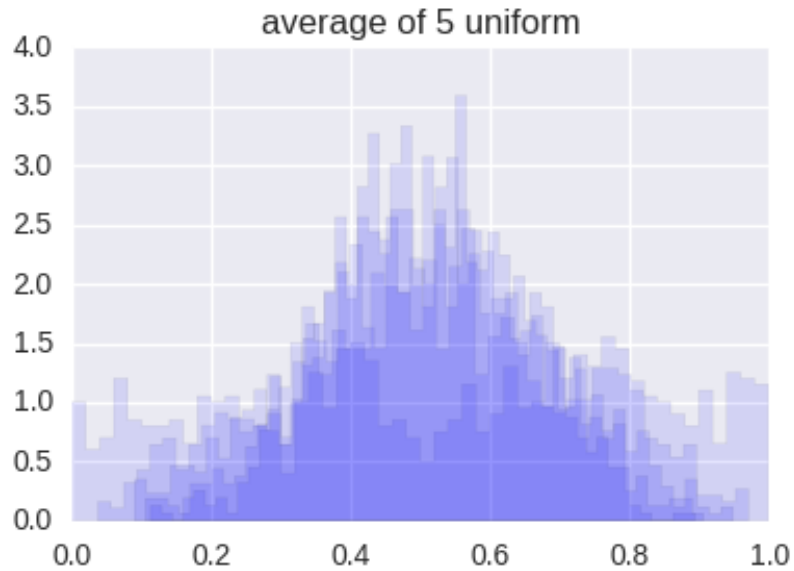


[1] **Bishop:2006ui**.

Why Gaussian: Central Limit Theorem

The distribution of a linear combination of a large number of independent, identically distributed (iid) variables will tend to normal, regardless of the underlying distribution.[1]

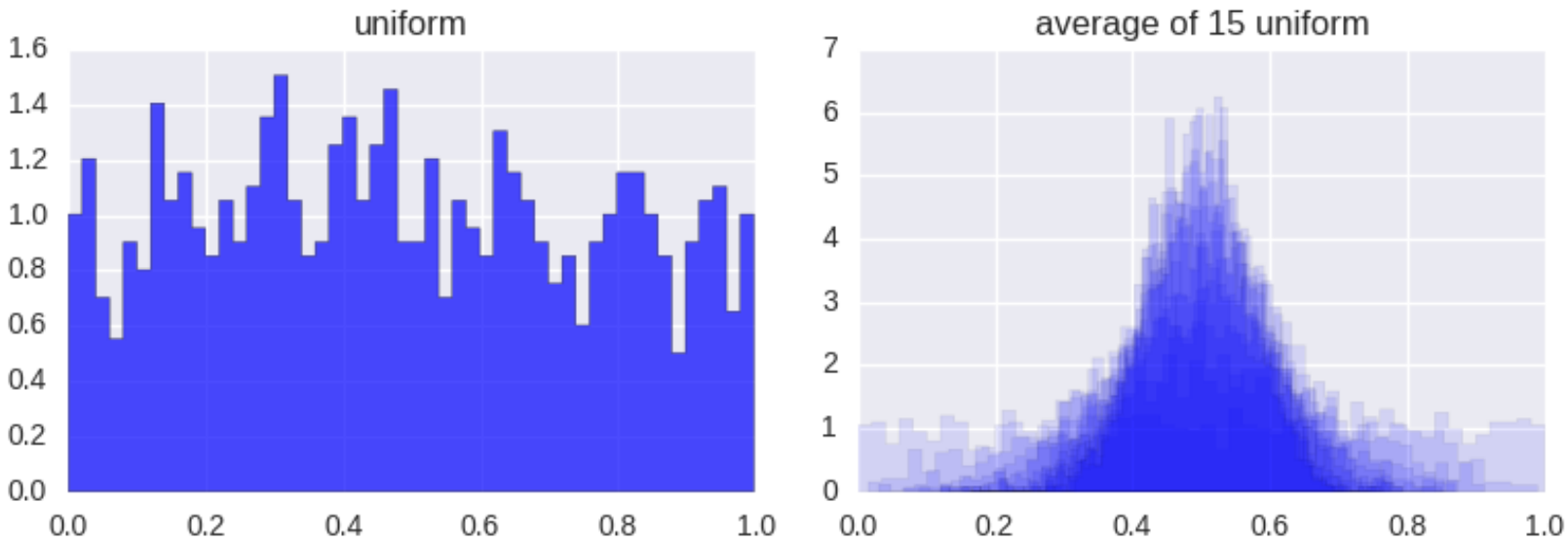




[1] **Bishop:2006ui**.

Why Gaussian: Central Limit Theorem

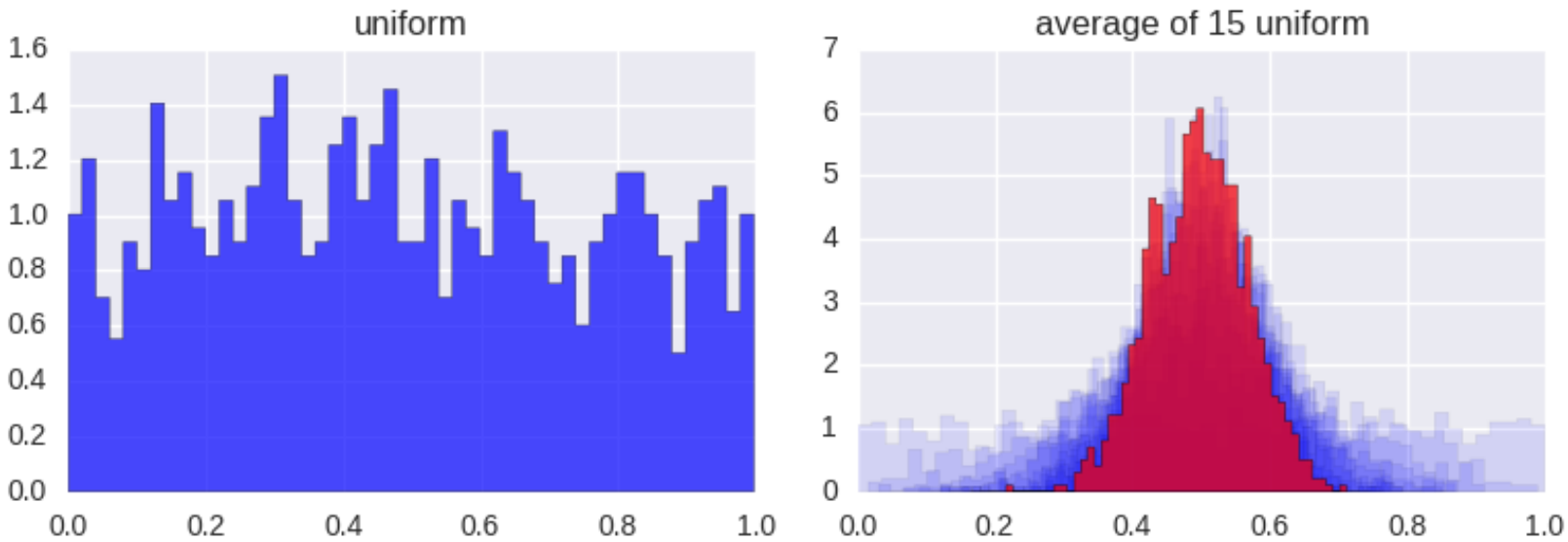
The distribution of a linear combination of a large number of independent, identically distributed (iid) variables will tend to normal, regardless of the underlying distribution.[1]



[1] **Bishop:2006ui**.

Why Gaussian: Central Limit Theorem

The distribution of a linear combination of a large number of independent, identically distributed (iid) variables will tend to normal, regardless of the underlying distribution.[1]



[1] **Bishop:2006ui**.

Why Gaussian: Central Limit Theorem

Galton Board (Sir Francis Galton, 1822-1911)

Gaussian distributions: D Dimensions

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

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

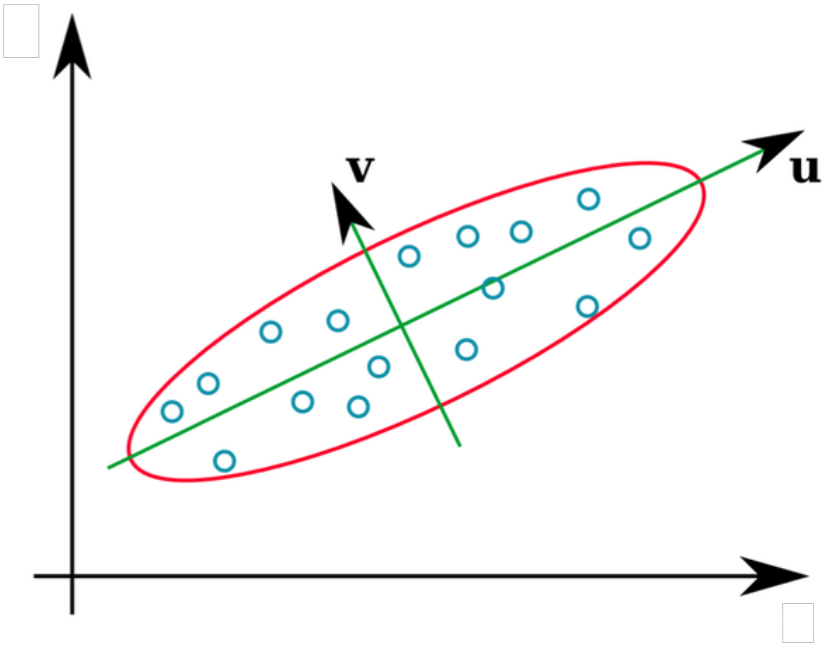
$$Pr(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

Gaussian distributions

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

Eigenvalue decomposition of the covariance matrix:

$$\Sigma = Q\ \Lambda\ Q^T$$

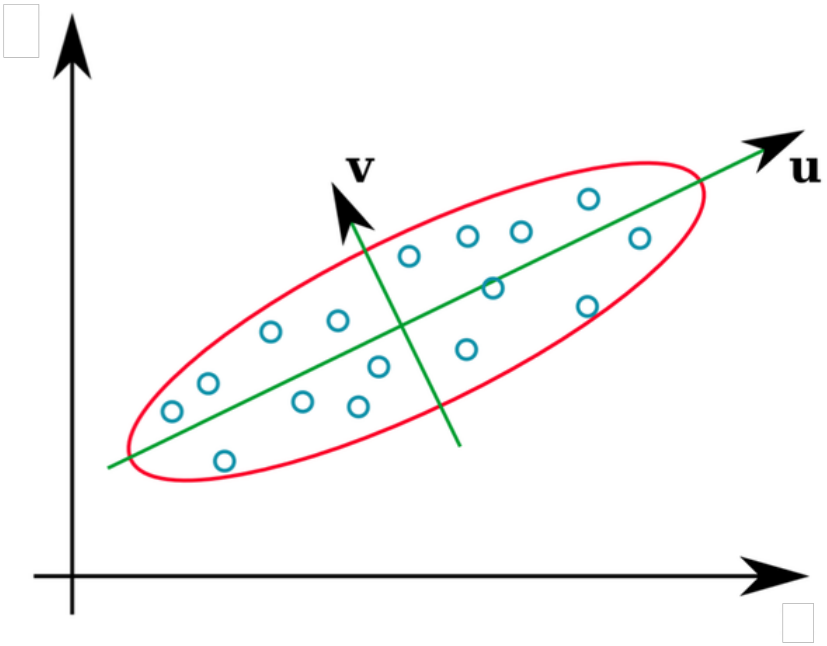


Gaussian distributions

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

Eigenvalue decomposition of the covariance matrix:

$$\Sigma = Q\ \Lambda\ Q^T$$

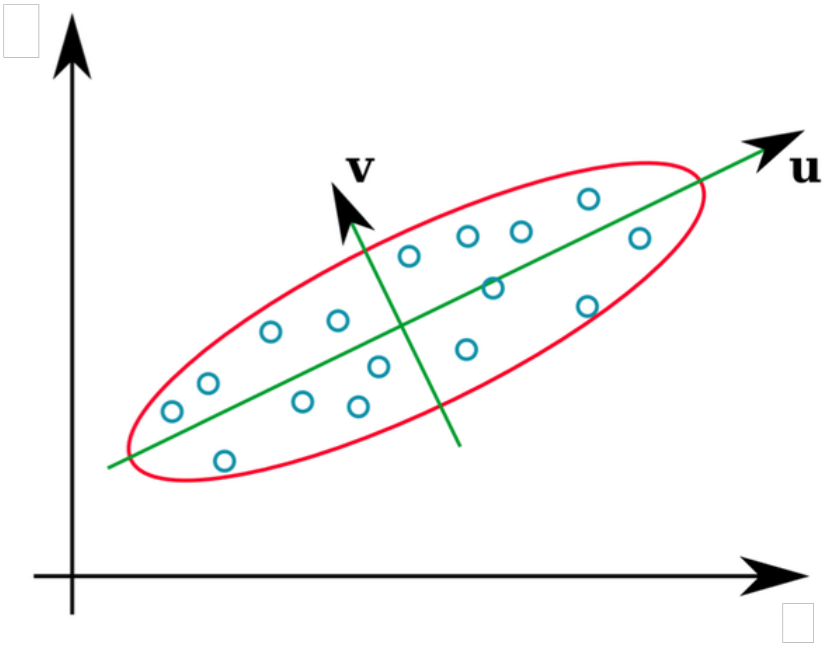


Gaussian distributions

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

Eigenvalue decomposition of the covariance matrix:

$$\Sigma = Q\ \Lambda\ Q^T$$



Expected value

$$\mathbb{E}[X] = \mu_X = \int x Pr(x) \mathrm{d}x$$

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

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

$$\overline{\mu_X} = \frac{1}{N} \sum_{i}^{N} x_i$$

Variance

$$\mathrm{Var}[X] = \sigma_X^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

- The "spread" of a distribution
- Unbiased sample variance

$$\overline{\sigma_X^2} = \frac{1}{N-1} \sum_i^N (x_i - \overline{\mu_X})^2$$

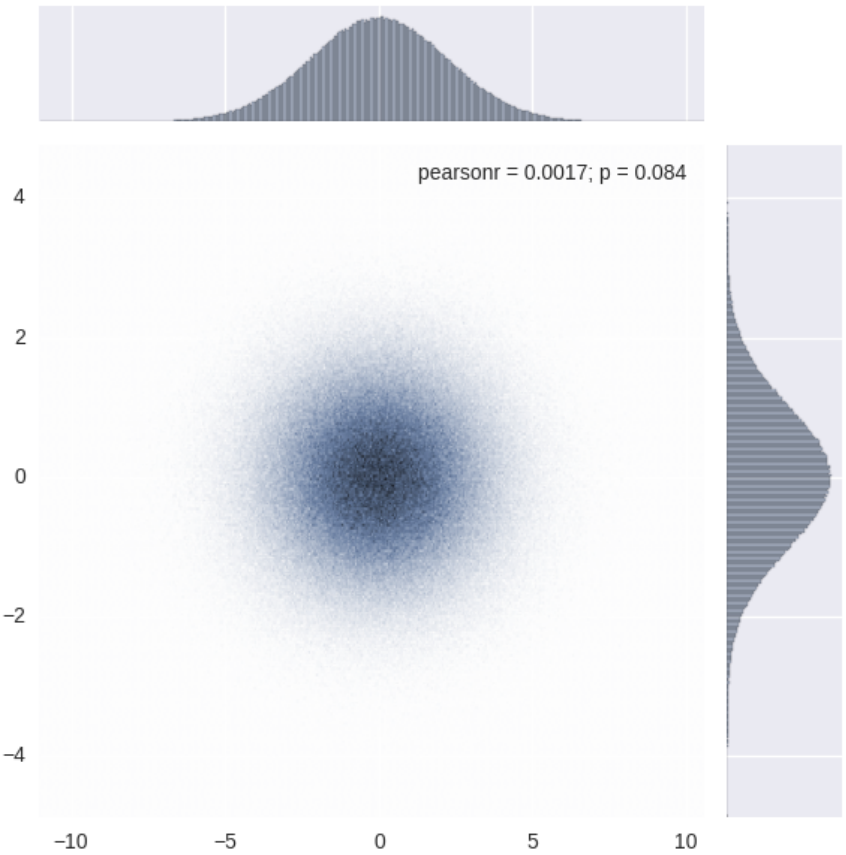
Covariance

$$\sigma_{X,Y} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$
$$\Sigma_{\mathbf{x}} = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}]))^T]$$

- Shows how two variables vary *together*
- Unbiased sample co-variance

$$\overline{\sigma_{X,Y}} = \frac{1}{N-1} \sum_{i}^{N} (x_i - \overline{\mu_X})(y_i - \overline{\mu_Y})$$

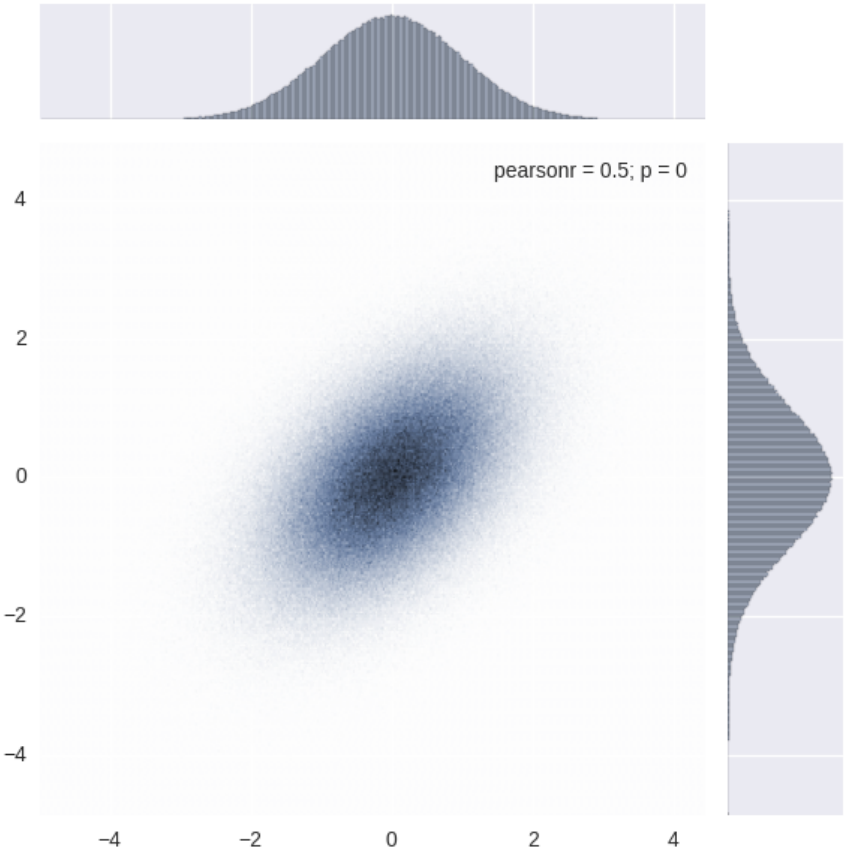
Examples



[2]

[2] Script by C.E. Ek

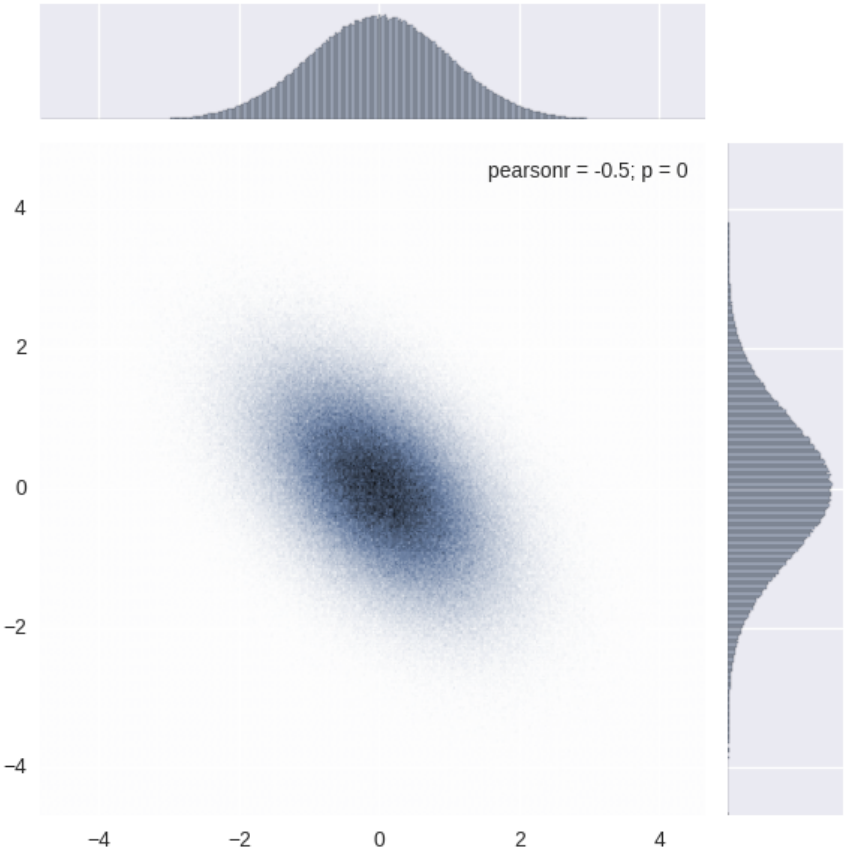
Examples



[2]

[2] Script by C.E. Ek

Examples

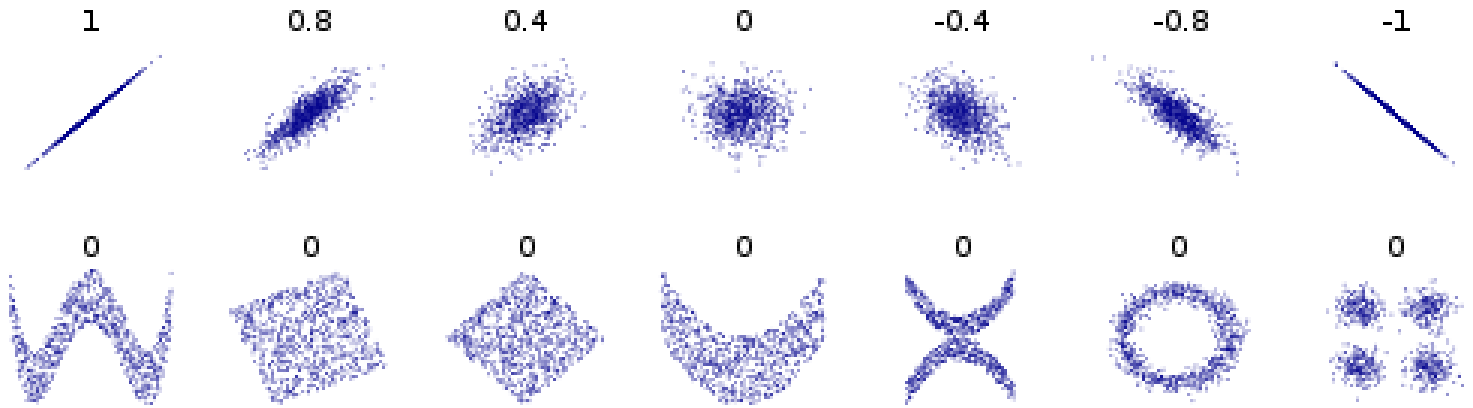


[2]

[2] Script by C.E. Ek

Covariance and Independence

- covariance shows "linear" dependency
- dependent variables may have zero covariance
- in some distributions zero covariance is equivalent to independence (example?)



3

[3] Figure adapted from Wikipedia

Outline

General ML problem (supervised learning)

Data:

$$\{(\mathbf{x}, y)_1, (\mathbf{x}, y)_2, \ldots, (\mathbf{x}, y)_n\}$$

Where $\mathbf{x}$ are features (independent variables), and y is the answer (dependent variable).

- if Y is discrete: *classification*
- if Y is continuous: *regression*

General ML problem (supervised learning)

Data:

$$\{(\mathbf{x}, y)_1, (\mathbf{x}, y)_2, \ldots, (\mathbf{x}, y)_n\}$$

Where $\mathbf{x}$ are features (independent variables), and y is the answer (dependent variable).

- if Y is discrete: *classification*
- if Y is continuous: *regression*

Learning: we want to learn how variables are related

- we estimate $Pr(\mathbf{x}, y)$ from observations $\{(\mathbf{x}, y)_j\}$

Inference: we want to predict an answer given an observation

- We estimate $Pr(y|\mathbf{X} = \mathbf{x})$ from observations $\{(\mathbf{x}, y)_j\}$

Bayes' Rule

$$Pr(y \,|\, \mathbf{X} = \mathbf{x}) = \frac{Pr(\mathbf{x} \,|\, Y = y) Pr(Y = y)}{Pr(\mathbf{X} = \mathbf{x})}$$

- $Pr(\mathbf{x} \,|\, Y = y) \leftarrow$ Likelihood represents the probability density of observing data $\mathbf{x}$ given the *hypothesis* $Y = y$.
- $Pr(Y = y) \leftarrow$ Prior represents the knowledge about Y before any observation.
- $Pr(y \,|\, \mathbf{X} = \mathbf{x}) \leftarrow$ Posterior represents the probability density of hypothesis y given observation $\mathbf{X} = \mathbf{x}$.
- $Pr(\mathbf{X} = \mathbf{x}) \leftarrow$ Evidence describes how well the model fits the evidence.

$$Pr(\mathbf{X} = \mathbf{x}) = \begin{cases} \sum_y Pr(\mathbf{x} \,|\, Y = y) Pr(Y = y) & \text{classification} \\ \int_y Pr(\mathbf{x} \,|\, Y = y) Pr(Y = y) & \text{regression} \end{cases}$$

Probabilistic Regression

Regression via conditional probability:

- Find joint distribution of $\mathbf{X}$ and Y: $Pr(\mathbf{x}, y)$
- Compute posterior of Y:
  $Pr(y|\mathbf{X} = \mathbf{x}) = Pr(\mathbf{x}, y)/Pr(\mathbf{X} = \mathbf{x})$
- Compute conditional expectation: $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$

Explicit regression model:

- Define a deterministic model $y = f(\mathbf{x}) + \epsilon$
- Describe probability distribution of the error $\epsilon = y - f(\mathbf{x})$
- Estimate parameters in $f(\mathbf{x})$

Example: Bivariate Normal distribution

Define joint probability distribution function

$$(X, Y) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Where:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

and ρ is the *correlation coefficient*. The posterior distribution is Normal:

$$Pr(y|X = x; \mu, \Sigma) = \mathcal{N}(\mu_{Y|X=x}, \sigma^2_{Y|X=x})$$

where:

$$\mu_{Y|X=x} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = w_0 + w_1 x$$

$$\sigma^2_{Y|X=x} = (1 - \rho^2)\sigma_Y^2 \quad \text{(constant wrt. } x\text{)}$$

Explicit Regression Model

Model (deterministic):

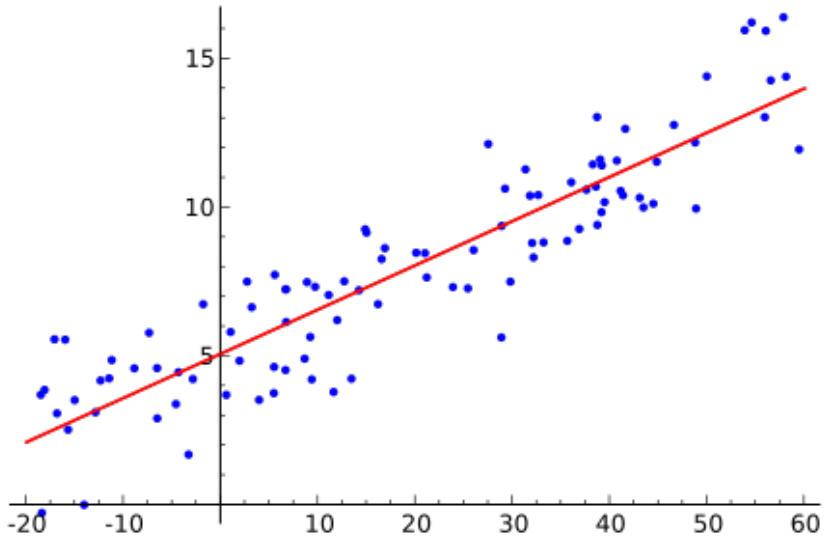
$$y = \mathbf{w}^T \mathbf{x} + \epsilon$$

But now:

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

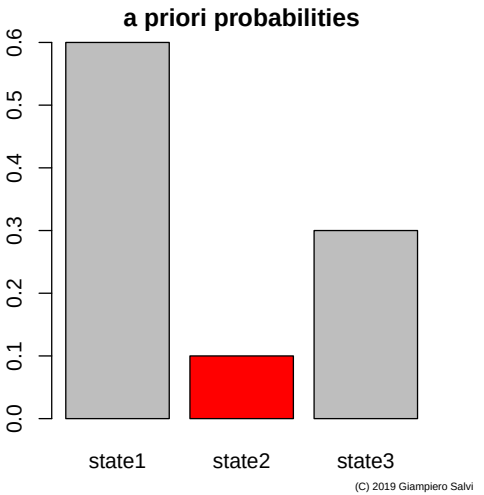
Therefore:

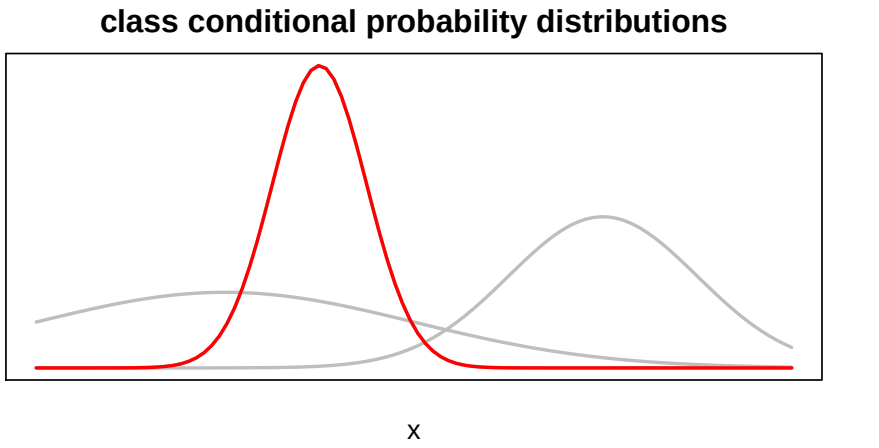
$$\begin{aligned} Y &\sim \mathcal{N}(\mu_Y(\mathbf{x}), \sigma_Y^2(\mathbf{x})) \\ &= \mathcal{N}(\mathbf{w}^T \mathbf{x}, \sigma^2) \end{aligned}$$



The Probabilistic Model of Classification

- One of K states is selected with *a priori* probability $Pr(Y = y)$
- Given state $Y = y$, an observation $\mathbf{X} \sim Pr(\mathbf{x}|Y = y)$





Problem

If I observe a new $\hat{\mathbf{x}}$, and I know $Pr(Y = y)$ and $Pr(\mathbf{x}|Y = y)$ for each of K classes, what can I say about $Pr(y|\mathbf{X} = \hat{\mathbf{x}})$?

Problem

If I observe a new $\hat{\mathbf{x}}$, and I know $Pr(Y = y)$ and $Pr(\mathbf{x}|Y = y)$ for each of K classes, what can I say about $Pr(y|\mathbf{X} = \hat{\mathbf{x}})$?

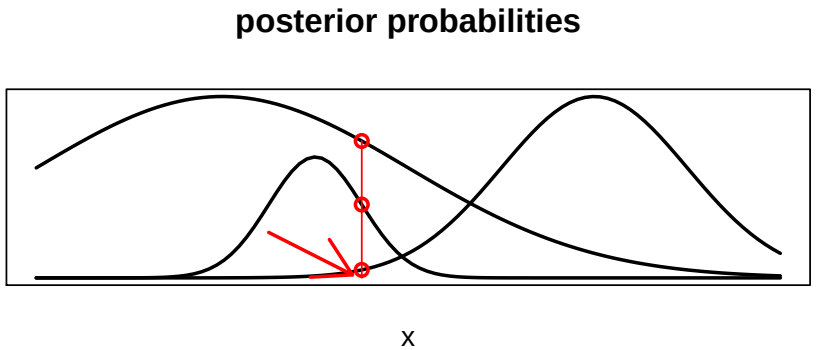
$$Pr(y|\mathbf{X} = \hat{\mathbf{x}}) = \frac{Pr(\mathbf{x}|Y = y)\ Pr(Y = y)}{Pr(\mathbf{X} = \hat{\mathbf{x}})}$$

Problem

If I observe a new $\hat{\mathbf{x}}$, and I know $Pr(Y = y)$ and $Pr(\mathbf{x}|Y = y)$ for each of K classes, what can I say about $Pr(y|\mathbf{X} = \hat{\mathbf{x}})$?

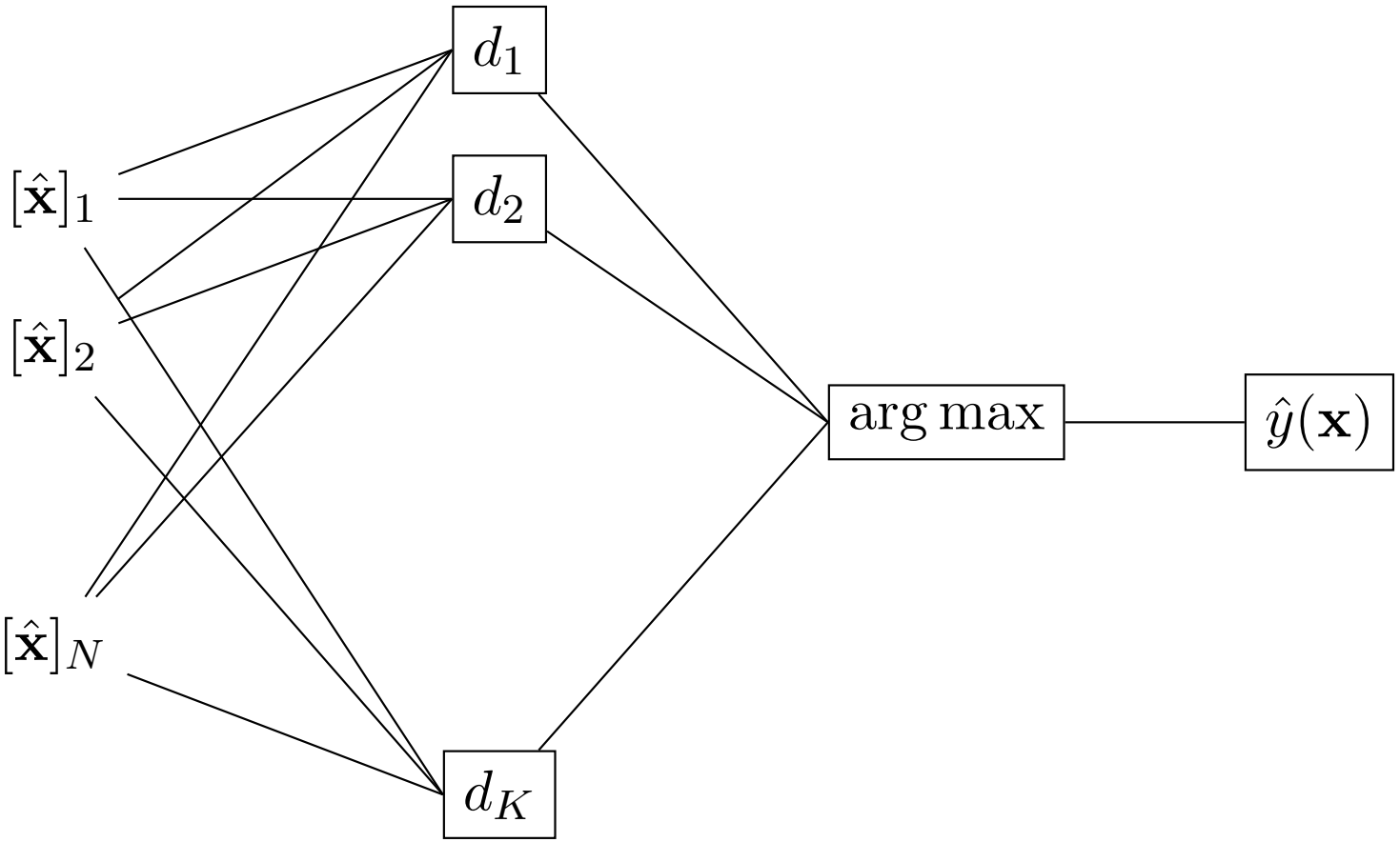
$$Pr(y|\mathbf{X} = \hat{\mathbf{x}}) = \frac{Pr(\mathbf{x}|Y = y)\ Pr(Y = y)}{Pr(\mathbf{X} = \hat{\mathbf{x}})}$$

$$\begin{aligned} Pr(\mathbf{X} = \hat{\mathbf{x}}) = & Pr(\mathbf{X} = \hat{\mathbf{x}}|Y = y_1)Pr(Y = y_1) \\ & + Pr(\mathbf{X} = \hat{\mathbf{x}}|Y = y_2)Pr(Y = y_2) \\ & + Pr(\mathbf{X} = \hat{\mathbf{x}}|Y = y_3)Pr(Y = y_3) \end{aligned}$$



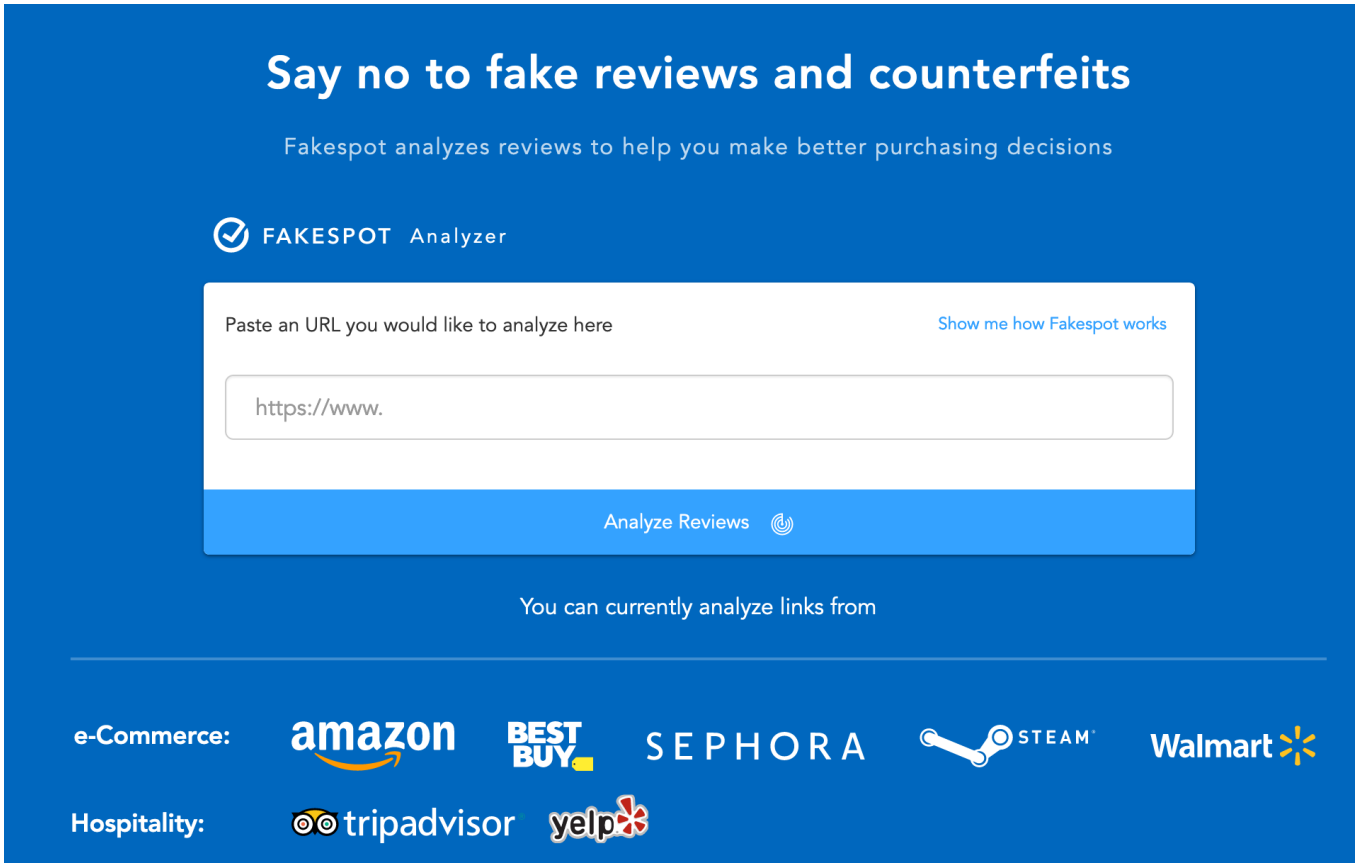
Classifiers: Discriminant Functions

$$d_k(\hat{\mathbf{x}}) = Pr(\hat{\mathbf{x}}|Y = y)\ P[Y = y]$$



Example: Is this online product review real or fake?

Task: Determine whether an online product review is real or fake given its length in characters.



Example: Is this online product review real or fake?

Task: Determine whether an online product review is real or fake given its length in characters.

Notation:

- Let $\Omega = \{\text{Real}, \text{Fake}\}$ be the sample space Ω
- Define r.v. $G : \{\text{Real}, \text{Fake}\} \to \{+1, -1\}$
- Let X be a r.v. mapping the length of a review in characters

Example: Is this online product review real or fake?

Task: Determine whether an online product review is real or fake given its length in characters.

Notation:

- Let $\Omega = \{\text{Real}, \text{Fake}\}$ be the sample space Ω
- Define r.v. $G : \{\text{Real}, \text{Fake}\} \to \{+1, -1\}$
- Let X be a r.v. mapping the length of a review in characters

Information given:

- Somehow, someone on the inside gave us a dataset of online product reviews labeled as real and fake. (Don't ask too many questions.)
- We can thus estimate the priors

$$P[\text{Real}] = P[G = +1] = 0.25, \quad P[\text{Fake}] = P[G = -1] = 0.75$$

- And we can estimate the likelihood distributions $Pr(x|G = +1)$ and $Pr(x|G = -1)$ (after we make a decision about the distribution)

Example: Is this online product review real or fake?

Task: Determine whether an online product review is real or fake given its length in characters. $\Longrightarrow$ Find $Pr(g|X = x)$.

Solution: Apply Bayes' Rule!

$$Pr(g|X = x) = \frac{Pr(X = x|G = g)P[G = g]}{Pr(X = x)}$$

$$= \frac{Pr(X = x|\, G = g)P[G = g]}{Pr(X = x|G = +1)P[G = +1] + P(X = x|G = -1)P[G = -1]}$$

We can now calculate $Pr(G = +1|X = x)$.

Selecting the most probable hypothesis

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

  Choose hypothesis from $\mathcal{Y}$ with highest probability given observed data $\mathbf{x}$:

$$\begin{aligned} y_{\text{MAP}}(\mathbf{x}) &= \arg\max_{y \in \mathcal{Y}} Pr(Y = y \,|\mathbf{X} = \mathbf{x}) \\ &= \arg\max_{y \in \mathcal{Y}} \frac{Pr(\mathbf{X} = \mathbf{x} \,|\, Y = y)\, Pr(Y = y)}{Pr(\mathbf{X} = \mathbf{x})} \\ &= \arg\max_{y \in \mathcal{Y}} Pr(\mathbf{X} = \mathbf{x} \,|\, Y = y)\, Pr(Y = y) \end{aligned}$$

Selecting the most probable hypothesis

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

  Choose hypothesis from $\mathcal{Y}$ with highest probability given observed data $\mathbf{x}$:

$$\begin{aligned} y_{\text{MAP}}(\mathbf{x}) &= \arg\max_{y \in \mathcal{Y}} Pr(Y = y \,|\mathbf{X} = \mathbf{x}) \\ &= \arg\max_{y \in \mathcal{Y}} \frac{Pr(\mathbf{X} = \mathbf{x} \,|\, Y = y)\, Pr(Y = y)}{Pr(\mathbf{X} = \mathbf{x})} \\ &= \arg\max_{y \in \mathcal{Y}} Pr(\mathbf{X} = \mathbf{x} \,|\, Y = y)\, Pr(Y = y) \end{aligned}$$

- **Maximum Likelihood (*ML*):**

  If we do not know prior distribution, then choose hypothesis with highest likelihood of generating the observed data:

$$y_{\text{MLE}} = \arg\max_{y \in \mathcal{Y}} Pr(\mathbf{X} = \mathbf{x} \,|\, Y = y)$$

Example: OMGROFL11! or Not?

Scenario: A patient has a biopsy to determine if they have Ontogenetic MytoGlenic Rennial Ocullar FibrioLyalgia Eleven! (OMGROFL11!). Of the entire population, only 0.8% have OMGROFL11!. The biopsy test returns a correct positive result in 98% of the cases in which OMGROFL11! is actually present, and a correct negative result in 97% of the cases in which it is not present. That's really accurate. The biopsy results comes back positive!!

Scenario as seen with probability:

Example: OMGROFL11! or Not?

Scenario: A patient has a biopsy to determine if they have Ontogenetic MytoGlenic Rennial Ocullar FibrioLyalgia Eleven! (OMGROFL11!). Of the entire population, only 0.8% have OMGROFL11!. The biopsy test returns a correct positive result in 98% of the cases in which OMGROFL11! is actually present, and a correct negative result in 97% of the cases in which it is not present. That's really accurate. The biopsy results comes back positive!!

Scenario as seen with probability:

- **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: OMGROFL11! or Not?

MAP:
When test returns a positive result,

$$y_{\text{MAP}} = \arg \max_{y \in \{\text{disease, not disease}\}} P[y \,|\, +]$$

$$= \arg \max_{y \in \{\text{disease, not disease}\}} P[+ \,|\, y]\, P[y]$$

Example: OMGROFL11! or Not?

MAP:
When test returns a positive result,

$$y_{\text{MAP}} = \arg \max_{y \in \{\text{disease, not disease}\}} P[y \,|\, +]$$
$$= \arg \max_{y \in \{\text{disease, not disease}\}} P[+ \,|\, y]\, P[y]$$

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

Example: OMGROFL11! or Not?

MAP:
When test returns a positive result,

$$y_{\text{MAP}} = \arg \max_{y \in \{\text{disease, not disease}\}} P[y \,|\, +]$$
$$= \arg \max_{y \in \{\text{disease, not disease}\}} P[+ \,|\, y]\, P[y]$$

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:

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

Summary

Today:

1. Probability Theory Reminder
   - Axioms and Properties
   - Common Distributions
   - Expectation
2. Probabilistic Machine Learning
   - Supervised Learning, General Definition
   - Regression
   - Classification

Next time: How to *fit* probability models to data (estimating parameters with the maximum likelihood optimality criterion)