Machine Learning

Lecture 3: probability & statistics refresher

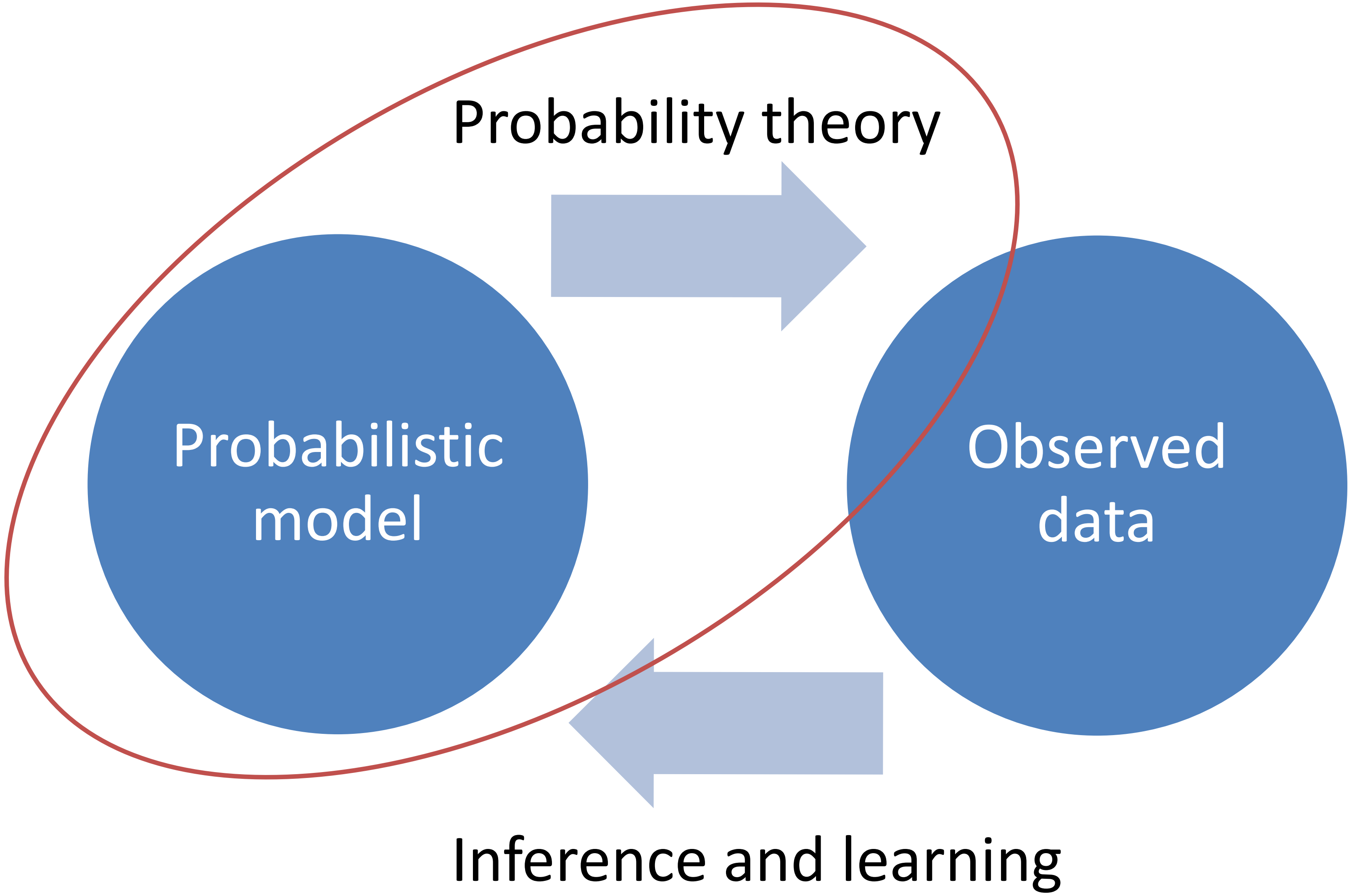
Jan Chorowski
Instytut Informatyki
Wydział Matematyki i Informatyki Uniwersytet Wrocławski
2020

Slides taken from presentations by: Y. LeCun, A. Ng, LXMLS Summer School

Additional materials

- http://cs229.stanford.edu/section/cs229-prob.pdf
- https://argmax.ai/docs/ml-course/01_lectureslides_ProbTheory.pdf
- Murphy, chapter 2
- Goodfellow et al. chapter 3 (the book webpage also hosts slides)
- Slides from LXMLS Summer School: http://lxmls.it.pt/2016/Lecture_0.pdf

Statistical modeling and inference



Definitions

- Ω is a **sample space**, e.g. two coin tosses $\Omega = \{HH, HT, TH, TT\}$
- $A \in 2^{\Omega}$ is an **event**, e.g. “first head” $\{HH, HT\}$

- $P: 2^{\Omega} \to \mathbb{R}$ is a **probability distributions** if:
  – $P(A) \geq 0$ for every A
  – $P(\Omega) = 1$
  – If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$

Discrete probability properties

- If $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(A \cap B) \leq \min\big(P(A), P(B)\big)$
- (Union bound) $P(A \cup B) \leq P(A) + P(B)$
- $P(\Omega \backslash A) = 1 - P(A)$
- (Law of Total Probability)
  If $A_1 \ldots A_k$ are disjoint and $\bigcup_{i=1}^{k} A_i = \Omega$, then
  $\sum_{i=1}^{k} P(A_i) = 1$.

Random Variables

A RV is a mapping $X\colon\ \Omega \rightarrow \mathbb{R}$.

- Discrete RV has countable values: $\{0,1\}$, $\mathbb{N}$
- RV X takes value x with a probability $\mathrm{P_X}(x = X)$
- E.g. Binomial distribution
  X is the number of heads in n tosses. Tosses are independent, each with head probability Θ.

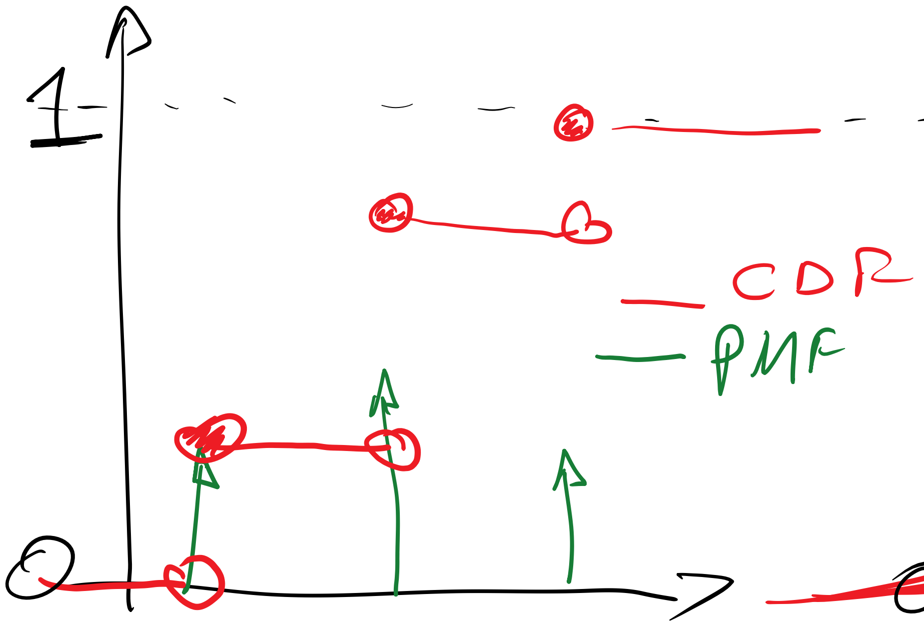
$$P_X(X = k) = P_X(k) = \binom{n}{k} \Theta^k (1 - \Theta)^{n-k}$$

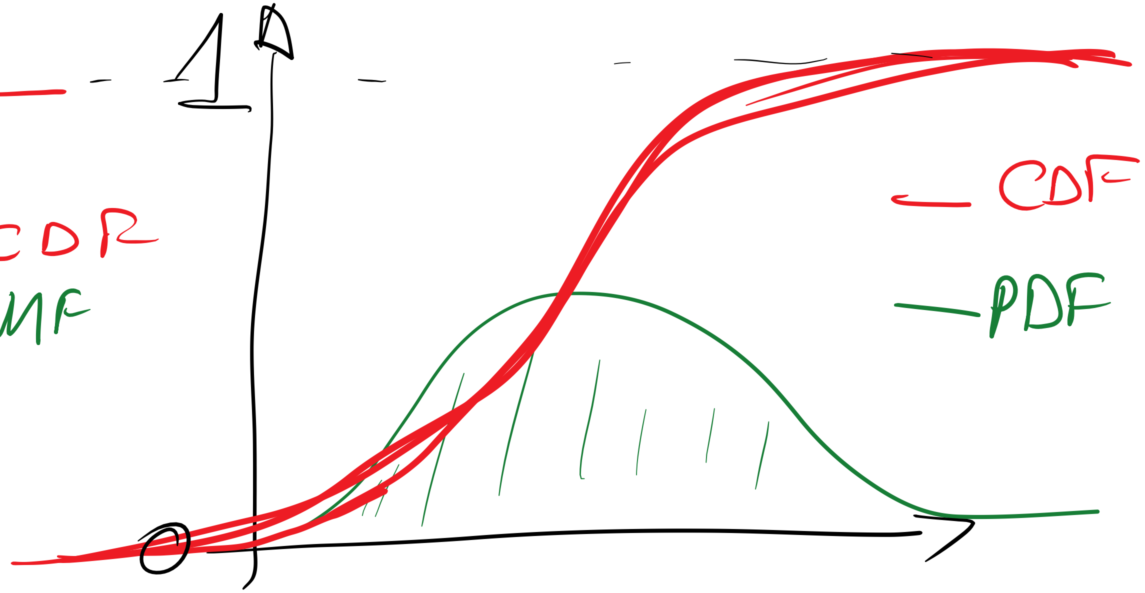
Continuous RV

- Continuous RV has uncountable values: $[0,1], \mathbb{R}$
- A continuous RV X has an associated **Probability Density Function** $f_X(x)$:
  - $\forall x f_X(x) \geq 0$
  - $\int_{-\infty}^{\infty} f_X(x)dx = 1$
  - $P(a < X \leq b) = \int_a^b f_X(x)dx$
  - For a continuous RV it is possible that $f_X(x) > 1$!
- Note: in the later lectures we will drop the distinction between probability $P()$ and probability density $f()$, using $P()$ in both contexts.

Cumulative distribution function (CDF)

- $F_X(x) = P_X(X \leq x)$
- $F_X(x) = \sum_{t \leq x} P_X(T)$ $\qquad F_X(x) = \int_{-\infty}^{x} f_X(t)dt$





Transformation of RVs

$$Y = g(X)$$

$$P_Y(y) = \sum_{x:y=g(x)} P_X(x)$$
$$= \sum_{x \in g^{-1}(y)} P_X(x)$$

$$f_Y(y) = f_X\left(g^{-1}(y)\right) \left|\frac{\partial g^{-1}(y)}{\partial y}\right|$$
$$= f_X(x) \left|\frac{\partial x}{\partial y}\right|$$

Assumption:
g is a bijection

Intuition:

$$f_Y(y)dy \approx f_X(x)dx$$

Expected values

- The expected value of a function r of a RV X is:

$$\mathbb{E}[r(X)]_{X\sim\, P(x)} = \sum_x r(x)P\,(x)$$

$$\mathbb{E}[r(X)]_{X\sim\, f_X} = \int r(x)f_X(x)dx$$

- Example: the mean value of X is $\mu = \sum_x xP(x)$
- The expectation is linear:
  - $\mathbb{E}[X + c] = \mathbb{E}[X] + c$ $\qquad\qquad$ $\mathbb{E}[cX] = c\mathbb{E}[X]$
  - $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ for all RV X and Y.

Variance

- Variance measures the spread of a RV X:

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

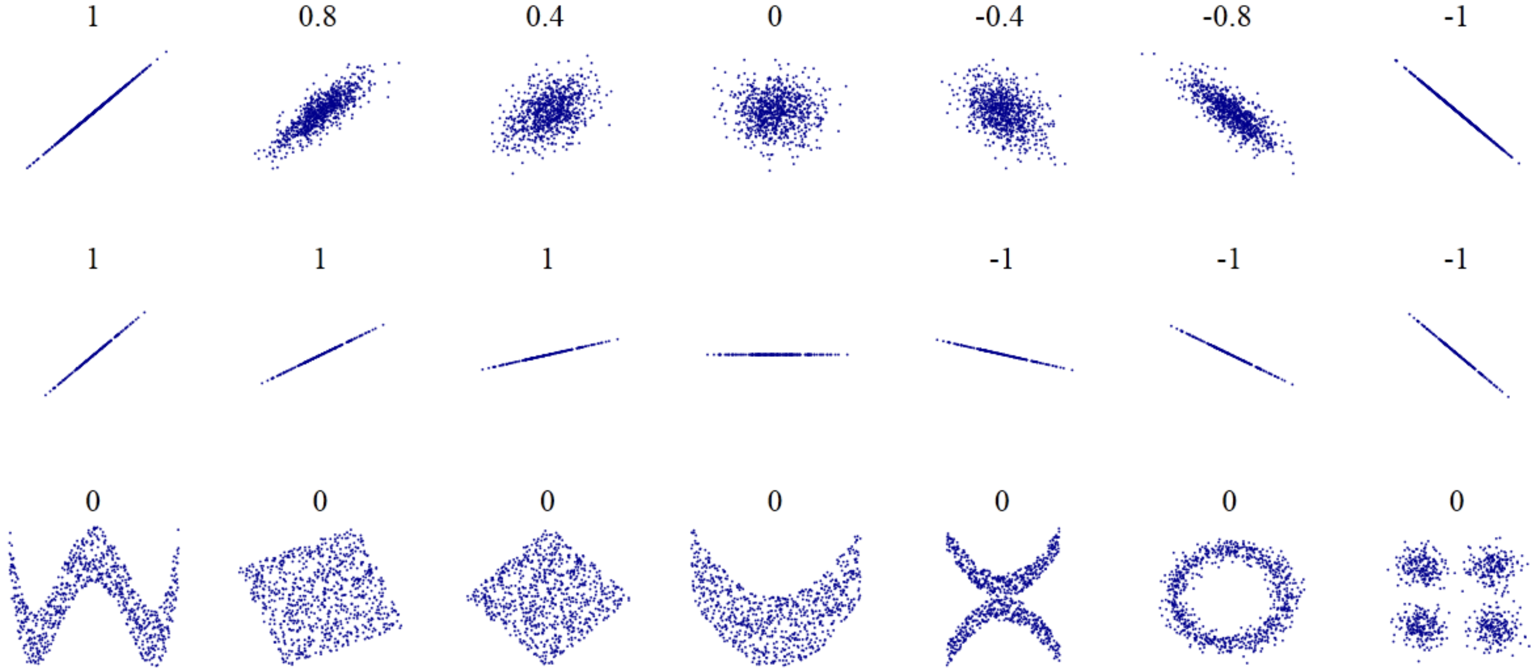
- Standard deviation $\sigma_X = \sqrt{\mathrm{Var}[X]}$
- The Covariance between X and Y is:
  $\mathrm{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$
- Properties of variance:
  - $\mathrm{Var}[X - c] = \mathrm{Var}[X]$
  - $\mathrm{Var[cX] = c^2Var[X]}$
  - $\mathrm{Var[aX + bY] = a^2Var[X] + b^2Var[Y]} + 2ab\mathrm{Cov[X, Y]}$
  - When X and Y are independent:
    $\mathrm{Var[aX + bY] = a^2Var[X] + b^2Var[Y]}$

Correlation

- Correlation coefficient is normalized Covariance:

$$\rho_{X,Y} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X \sigma_Y}$$

- $-1 \le \rho_{X,Y} \le 1$
- Independent $\Rightarrow$ uncorrelated



Joint probability

- Given two RVs X and Y $P(x, y)$ denotes the event that $X = x$ and $Y = y$.
- X and Y are independent iff $P(x, y) = P(x)P(y)$
- Marginal probability: $P(x) = \sum_y P(x, y)$
- Conditional probability (read probability of x given y):

$$P(x|y) = \frac{P(x, y)}{P(y)}$$

Bayes theorem

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

$$\underbrace{P(x|y)}_{\text{posterior}} = \frac{\overbrace{P(y|x)}^{\text{likelihood}}\overbrace{P(x)}^{\text{prior}}}{P(y)} = \frac{P(y|x)P(x)}{\sum_{x'} P(x',y)}$$

E.g. compute p(car crash | drunk driving)

Bayes theorem in action

We want to measure: $P(\mathrm{crash}|\mathrm{drunk})$

Can't get people drunk and send on the road…

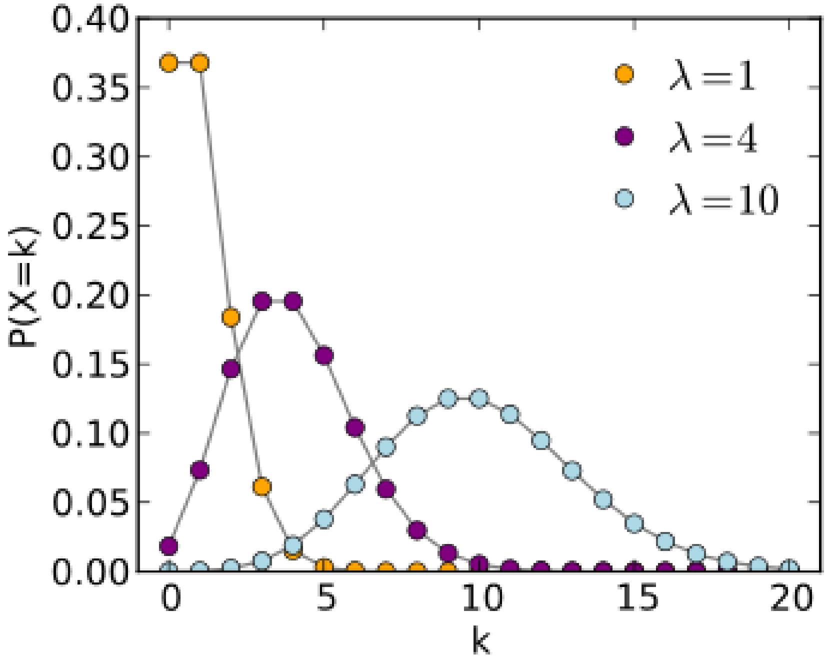
$$P(\mathrm{crash}|\mathrm{drunk}) = \frac{P(\mathrm{drunk}|\mathrm{crash})P(\mathrm{crash})}{P(\mathrm{drunk})}$$

That's ethical – we can estimate all need probabilities from police statistics!

Bernoulli and Binomial

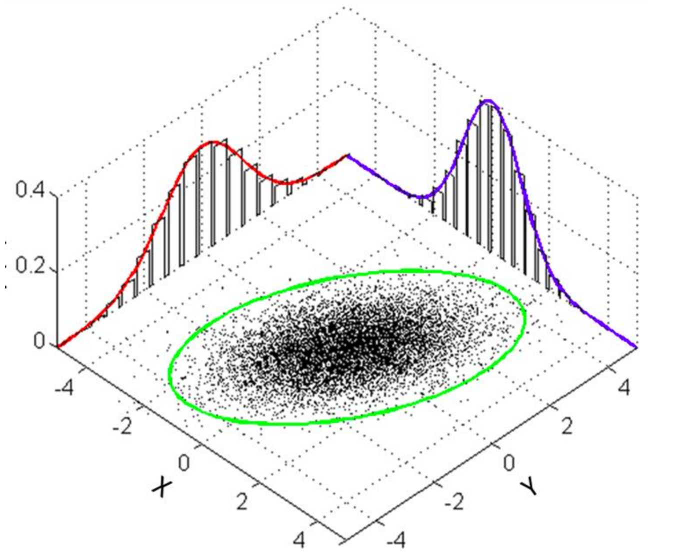
- Bernoulli:
  - X is binary
    $P(X = 1) = \phi, P(X = 0) = 1 - \phi$
  - $\mathbb{E}[X] = 0(1 - \phi) + 1\phi = \phi$
  - $\mathrm{Var}[X] = (0 - \phi)^2(1 - \phi) + (1 - \phi)^2\phi = \phi(1 - \phi)$
- Binomial:
  - RV K = sum of n independent Bernoulli(ϕ) trials
  - $P(k; \phi, n) = \binom{n}{k} \phi^k(1 - \phi)^{n-k}$
  - $\mathbb{E}[K] = n\phi$
  - $\mathrm{Var}(K) = n\phi(1 - \phi)$

Poisson



- The count of rare events
- Defined for natural numbers
- $P(X = k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$
- $\mathbb{E}[X] = \lambda$
- $\mathrm{Var}[X] = \lambda$
- Sum of independent Poissons is Poisson: if $X \sim \mathrm{Pois}(\lambda_X)$ and $Y \sim \mathrm{Pois}(\lambda_Y)$ then $X + Y \sim \mathrm{Pois}(\lambda_X + \lambda_Y)$

Normal distribution



- $X \sim \mathcal{N}(\mu, \sigma^2)$
- Univariate:

$$P(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Multivariate, k-dimensional:

$$P(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$

- Mean: $\boldsymbol{\mu}$
- Variance: $\boldsymbol{\Sigma}$ (in 1D case σ)
- Conditionals, sums, and marginals of Gaussians are Gaussian