Probabilistic Machine Learning:
2. Statistical refresher

Tomasz Kajdanowicz, Przemysław Kazienko (zastępstwo)

Department of Computational Intelligence
Wroclaw University of Science and Technology

The presentation has been inspired and in some parts totally based on Prof. Mario A. T. Figueiredo presentation at LxMLS'2017, Instituto Superior Tecnico & Instituto de Telecomunicacoes, Lisboa, Portugal.
Apropriate agreements to propagate his ideas has been acquired.

Expectation of (Real) Random Variables

- Expectation:

$$\mathbb{E}(X) = \begin{cases} \sum_i x_i f_X(x_i) & \text{if } X \in \{x_1, \dots, x_k\} \subset \mathbb{R} \\ \int_{-\infty}^{\infty} x f_X(x) dx & \text{if } X \quad \text{continuous} \end{cases}$$

Example

- Bernoulli, $f_X(x) = p^x(1-p)^{1-x}$, for $x \in \{0, 1\}$
  $\mathbb{E}(X) = 0(1-p) + 1p = p$
- Binomial, $f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$, for $x \in \{0, \dots, n\}$
  $\mathbb{E}(X) = np$
- Gaussian, $f_X(x) = \mathcal{N}(x; \mu, \sigma^2)$
  $\mathbb{E}(X) = \mu$

- Linearity of expectation: $\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}(X) + \beta \mathbb{E}(Y), \quad \alpha, \beta \in \mathbb{R}$
  $\mathbb{E}(c) = c$

Expectation of Functions of RVs

- $$\mathbb{E}(g(X)) = \begin{cases} \sum_i g(x_i) f_X(x_i) & \text{if } X \quad \text{discrete, } g(x_i) \in \mathbb{R} \\ \int_{-\infty}^{\infty} g(x) f_X(x) dx & \text{if } X \quad \text{continuous} \end{cases}$$

Example

- variance, $var(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$
- Bernoulli variance, $\mathbb{E}(X^2) = \mathbb{E}(X) = p$, thus $var(X) = p(1 - p)$
- Gaussian variance, $\mathbb{E}((X - \mu)^2) = \sigma^2$

The importance of the Gaussian

The importance of the Gaussian

Take n independent RVs $X_1, \ldots, X_n$, with $\mathbb{E}[X_i] = \mu_i$ and $var(X_i) = \sigma_i^2$

- Their sum, $Y_n = \sum_{i=1}^{n} X_i$ satisfies:

$$\mathbb{E}[Y_i] = \sum_{i=1}^{n} \mu_i \equiv \mu$$

$$var(Y_n) = \sum_{i=1}^{n} \sigma_i^2 \equiv \sigma^2$$

- Let $Z_n = \frac{Y_n - \mu}{\sigma}$, thus $\mathbb{E}[Z_n] = 0$ and $var(Z_n) = 1$
- Central limit theorem: under mild conditions:

$$lim_{n \to \infty} Z_n \sim \mathcal{N}(0, 1)$$

Two (or More) Random Variables

- Joint pmf of two discrete RVs: $f_{X,Y}(x,y) = \mathbb{P}(X = x \wedge Y = y)$
  Extends trivially to more than two RVs.
- Joint pdf of two continuous RVs: $f_{X,Y}(x,y)$, such that

$$\mathbb{P}((X,Y) \in A) = \int\int_A f_{X,Y}(x,y)dxdy, \quad A \in \sigma(\mathbb{R}^2)$$

  Extends trivially to more than two RVs.
- Marginalization:

$$f_Y(y) = \begin{cases} \sum_x f_{X,Y}(x,y) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx & \text{if } X \text{ continuous} \end{cases}$$

- Independence: $(X \perp\!\!\!\perp Y) \iff f_{X,Y} = f_X(x)f_Y(y) \implies \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$

Conditionals and Bayes' Theorem

- Conditional pmf:

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

- Conditional pdf:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

...the meaning is technically delicate

- Bayes' theorem:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}$$

- Also valid in the mixed case (e.g., X continuous, Y discrete)

Joint, Marginal, and Conditional Probabilities: An Example

- A pair of binary variables $X, Y \in \{0, 1\}$, with joint pmf:

$f_{X,Y}(x, y)$	$Y = 0$	$Y = 1$
$X = 0$	1/5	2/5
$X = 1$	1/10	3/10

Joint, Marginal, and Conditional Probabilities: An Example

- A pair of binary variables $X, Y \in \{0, 1\}$, with joint pmf:

$f_{X,Y}(x,y)$	$Y = 0$	$Y = 1$
$X = 0$	1/5	2/5
$X = 1$	1/10	3/10

- Marginals: $f_X(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5}$
  $f_X(1) = \frac{1}{10} + \frac{3}{10} = \frac{4}{10}$
  $f_Y(0) = \frac{1}{5} + \frac{1}{10} = \frac{3}{10}$
  $f_Y(1) = \frac{2}{5} + \frac{3}{10} = \frac{7}{10}$
- Conditional probabilities:

Joint, Marginal, and Conditional Probabilities: An Example

- A pair of binary variables $X, Y \in \{0, 1\}$, with joint pmf:

$f_{X,Y}(x,y)$	$Y=0$	$Y=1$
$X=0$	1/5	2/5
$X=1$	1/10	3/10

- Marginals: $f_X(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5}$
  $f_X(1) = \frac{1}{10} + \frac{3}{10} = \frac{4}{10}$
  $f_Y(0) = \frac{1}{5} + \frac{1}{10} = \frac{3}{10}$
  $f_Y(1) = \frac{2}{5} + \frac{3}{10} = \frac{7}{10}$
- Conditional probabilities:

$f_{X\|Y}(x\|y)$	$Y=0$	$Y=1$
$X=0$	2/3	4/7
$X=1$	1/3	3/7

$f_{Y\|X}(y\|x)$	$Y=0$	$Y=1$
$X=0$	1/3	2/3
$X=1$	1/4	3/4

An Important Multivariate RV: Multinomial

- Multinomial: $X = (X_1, \ldots, X_K)$, $X_i \in \{0, \ldots, n\}$, such that $\sum_i X_i = n$, X_i denotes no. of outcomes class i occurs in n trials

$$f_X(x_1, \ldots, x_k) = \begin{cases} \binom{n}{x_1 x_2 \ldots x_K} p_1^{x_1} p_2^{x_2} \ldots p_K^{x_K} & \text{if } \sum_{i=1}^K x_i = n \\ 0 & \text{if } \sum_{i=1}^K x_i \neq n \end{cases}$$

$$\binom{n}{x_1 x_2 \ldots x_K} = \frac{n!}{x_1! x_2! \ldots x_K!}$$

- Generalizes the binomial from binary to K-classes.

Example

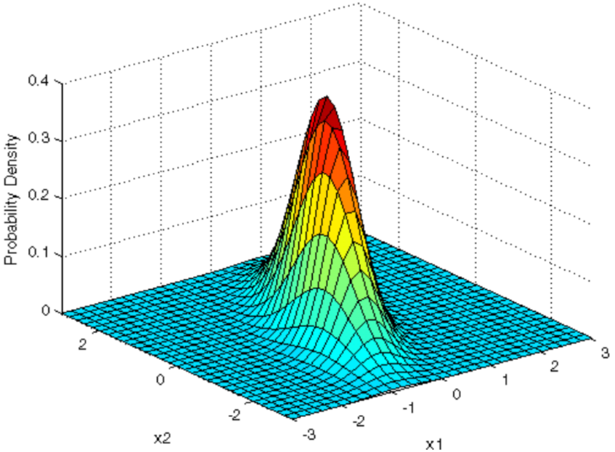
- tossing n independent fair dice, $K = 6$, $p_1 = \ldots = p_6 = 1/6$, x_i = number of outcomes with i dots (of course, $\sum_{i=1}^{K=6} x_i = n$)
- bag of words (BoW) multinomial model with vocabulary of K words and the document that consists of n words in total.

An Important Multivariate RV: Gaussian

- Multivariate Gaussian: $X \in \mathbb{R}^n$

$$f_X(x) = \mathcal{N}(x; \mu, C) = \frac{1}{\sqrt{det(2\pi C)}} exp\Big(-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\Big)$$

- Parameters: vector $\mu \in \mathbb{R}^n$ and matrix $C \in \mathbb{R}^{n\times n}$
- Expected value: $\mathbb{E}(X) = \mu$. Meaning of C: next slide



Covariance, Correlation, and all that...

- Covariance between two RVs:

$$cov(X,Y) = \mathbb{E}\Big[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\Big] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

- Relationship with variance: $var(X) = cov(X,X)$
- Correlation:

$$corr(X,Y) = \rho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)}\sqrt{var(Y))}} \in [-1,1]$$

- $X \perp\!\!\!\perp Y \iff f_{X,Y}(x,y) = f_X(x)f_Y(y) \implies cov(X,Y) = 0$

More on Covariances

Let X, Y, W, V are real-valued random variables and a, b, c, d are constant.

- $cov(X, a) = 0$
- $cov(X, X) = var(X)$
- $cov(X, Y) = cov(Y, X)$
- $cov(aX, bY) = abcov(X, Y)$
- $cov(X + a, Y + b) = abcov(X, Y)$
- $cov(aX + bY, cW + dV) = accov(X, W) + adcov(X, V) + bccov(Y, W) + bdcov(Y, V)$