A Refresher on Probabilities

Pratanu Roy and Alexey (Alex) Gronskiy

September 24 – 26, 2014

Tutorial Outline

1. Preliminaries
2. A Refresher on Probabilities
3. Convergence of RVs

Literature

- **Christopher M. Bishop, Pattern Recognition and Machine Learning. Springer Verlag (2006)**
- Richard O. Duda, Peter E. Hart & David G. Stork, *Pattern Classification*. Wiley & Sons (2001)
- Trevor Hastie, Robert Tibshirani & Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Verlag (2001)
- Luc Devroye, Laslo Györfi & Gabor Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer Verlag (1996)
- Vladimir N. Vapnik, *Estimation of Dependences Based on Empirical Data*. Springer Verlag (1983)
- Ulf Grenander, General Pattern Theory: A Mathematical Study of Regular Structures. Oxford University Press (1993)
- Andrew Webb, *Statistical Pattern Recognition*. Wiley & Sons, (2002)
- Keinosuke Fukunaga, *Statistical Pattern Recognition*. Academic Press (1990)
- Brian D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge University Press (1996)
- Larry Wasserman, *All of Statistics*. (1st ed. 2004. Corr. 2nd printing, ISBN: 0-387-40272-1) Springer Verlag (2004)

Probability

- Fraction of times an event occurs.
- Degree of belief about an event.
- Useful as data model, incorporate noise, uncertainty, ...

Random Variables

- A random variable is a "probabilistic" outcome of an experiment, such as a coin flip or the height of a person chosen from a population.
- Notation:

X Random variable
$\approx$ a device from which we draw a value.

x If x is not capital, it denotes a realization of X.
$\Pr\{X = x\}$ denotes the probability for this to occur.

$\mathcal{X}$ Sample space or domain of X.
The set of all values a draw from X may result in.

Random Variables

- RVs take on values in a sample space $\mathcal{X}$. This space may be discrete or continuous, and the space may be defined differently for different scenarios.

Types of sample spaces:

1. Discrete sets:
    - Finite: for a coin flip $\mathcal{X} = \{H, T\}$
    - Infinite: $\mathcal{X} = \mathbb{N}, \mathbb{Z}$ etc.
2. Continuous sets: e.g. $\mathcal{X} = \mathbb{R}, \mathbb{R}_+, \mathbb{R}^d, [0, 1], [a, b]$

- There is not necessarily one uniquely “correct“ sample space for a particular concept.

Probability of Random Variables

Probability distribution function describes how probabilities are distributed over the values of the random variable.

$$\text{Probability(event)} = \frac{\text{The total number of events of interest}}{\text{The total number of events}}$$

Probability of Random Variables

- A discrete distribution assigns a probability to every atom in the sample space of a random variable.
- For example, if X is an (unfair) coin, then the sample space consists of the atomic events $X = H$ and $X = T$, and the discrete distribution might look like:
  $\Pr\{X = H\} = 0.7$
  $\Pr\{X = T\} = 0.3$
- For any valid discrete distribution, the probabilities over the atomic events must fulfill:
  1. Non-negativity: $\Pr\{x\} \geq 0$
  2. Normalization: $\sum_{x \in \mathcal{X}} \Pr\{X = x\} = 1$

Probability of Random Variables

- An event is a subset of atoms (one or more). The probability of an event is the sum of the probabilities of its constituent atoms.
- **Example:** Consider the event of a single die roll (D) is bigger than 3

  The probability of $D > 3$ is equivalent to the probability that the outcome is 4, or the outcome is 5, or the outcome is 6. The probabilities that the die is 4, 5, or 6 are added together:

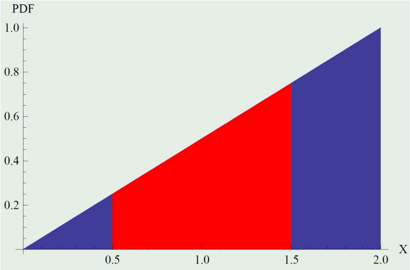
$$\mathrm{Pr}\{D > 3\} = \mathrm{Pr}\{D = 4\} + \mathrm{Pr}\{D = 5\} + \mathrm{Pr}\{D = 6\}$$

Continuous Random Variables

- A continuous random variable can assume any value in an interval or in a collection of intervals.

$$\text{Pr}\{a \leq X \leq b\} = \int_a^b p(x)dx$$

Example: Find the probability that $0.5 \leq X \leq 1.5$



Continuous Random Variables

- For continuous probability distributions, we require:
  1. Non-negativity: $p(x) \geq 0$
  2. Normalization: $\int_{\mathcal{X}} p(x)dx = 1$

- **Notation:** We deal with three types of symbols:

Pr$\{\ldots\}$	Probability of an event (inside the curly brackets), such as Pr$\{X = x\}$.
$P(x)$	Probability mass function.
$p(x)$	Probability density function.

- Density functions are only applicable in the case of continuous sample spaces.

Joint Probabilities

Typically, one considers collections of RVs.
For example, the flipping of 4 coins involves 4 RVs, 1 for each coin.

Joint probability:	The probability for precisely the values x, y to occur together.
Definition:	$P(x, y) := \Pr\{X = x, Y = y\}$

The joint distribution for a flip of each of 4 coins assigns a probability to every outcome in the space of all possible outcomes of the 4 flips.

If all coins are fair:

$$
\begin{aligned}
P(HHHH) &= 0.0625 \\
P(HHHT) &= 0.0625 \\
P(HHTH) &= 0.0625 \\
&\ldots
\end{aligned}
$$

Conditional Probability

A conditional distribution is the distribution of some random variable given some evidence, such as the value of another random variable.

- **Def.**: $\Pr\{X = x | Y = y\}$ is the probability that $X = x$ when $Y = y$.

A conditional distribution gives more information about X than the distribution of $P(X)$ alone.

Conditional Probability

The conditional distribution $\Pr\{X = x | Y = y\}$ is a different distribution for each value of y. Such that

$$\sum_x \Pr\{X = x | Y = y\} = 1$$

However, remember that, generally

$$\sum_y \Pr\{X = x | Y = y\} \neq 1.$$

Marginalization

- Given a collection of random variables, we are often interested in only a subset of them. For example, we might want to compute $P(X)$ from a joint distribution $P(X, Y, Z)$.

Def.

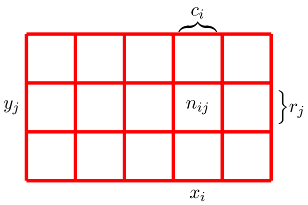
Marginal probability:	The probability for x to occur, regardless of y.
Discrete case:	$P(x) := \sum_{y \in \mathcal{Y}} P(x, y)$
Continuous case:	$p(x) := \int_{\mathcal{Y}} p(x, y) dy$

Marginalization

This property actually derives from the chain rule:

$$\begin{aligned}
\sum_{y\in\mathcal{Y}} P(x,y) &= \textstyle\sum_{y\in\mathcal{Y}} P(x)P(y|x) && \text{by the chain rule} \\
&= P(x)\textstyle\sum_{y\in\mathcal{Y}} P(y|x) && P(x) \text{ doesn't depend on y} \\
&= P(x) && \textstyle\sum_{y\in\mathcal{Y}} P(y|x) = 1
\end{aligned}$$

Conditional, Joint, Marginal



Joint Probability

The entry of both values jointly.

$$\Pr\{X = x_i, Y = y_j\} = \frac{n_{ij}}{N}$$

Marginal Probability

The sum over a row or column.

$$\Pr\{X = x_i\} = \frac{c_i}{N}$$

Conditional Probability

The fraction of a row or column in a particular cell.

$$\Pr\{Y = y_j \mid X = x_i\} = \frac{n_{ij}}{c_i}$$

Simpson's Paradox

- illustrates the difference between marginal and conditional distributions
- men and women under treatment $\longrightarrow$ recovery?
- marginal of frequencies:

Treatment \ Recovery	0	1
0	180	180
1	200	200

obviously:

R indep. of T

- conditional tables of frequencies (given the gender):

Men:

Treatment \ Recovery	0	1
0	120	160
1	50	100

Women:

Treatment \ Recovery	0	1
0	60	20
1	150	100

obviously:

R dep. on T given M/F

The other way around

- men and women under treatment $\longrightarrow$ recovery?
- marginal table of frequencies:

Ttreatment	Recovery 0	Recovery 1
0	50	40
1	90	120

obviously:

R dep. on T

- conditional tables of frequencies (given the gender):

Men:

Ttreatment	Recovery 0	Recovery 1
0	40	20
1	40	20

Women:

Ttreatment	Recovery 0	Recovery 1
0	10	20
1	50	100

obviously:

R indep. of T given M/F

Conditional Probability and Related Concepts

Conditional probability can be defined in terms of the joint and single probability distributions:

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)}$$

(provided $P(Y) > 0$)

The Chain Rule

The definition of conditional probability leads to the chain rule, which lets us define the joint distribution of two (or more) random variables as a product of conditionals:

The Chain Rule:

$$\begin{aligned} P(X,Y) &= \frac{P(X,Y)P(Y)}{P(Y)} \\ &= P(X|Y)P(Y) \end{aligned}$$

- The chain rule can be used to derive the $P(X,Y)$ when it is not known.
- The chain rule can be extended to any set of n variables.

Bayes Rule

By the chain rule:

$$
\begin{aligned}
P(X, Y) &= P(X|Y)P(Y) \\
&= P(Y|X)P(X)
\end{aligned}
$$

This is equivalently expressed as Bayes rule:

$$
P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}
$$

posterior $\propto$ likelihood $\times$ prior

Independence

- Random variables are independent if knowing about X tells us nothing about Y. That is,

$$P(Y|X) = P(Y)$$

.

- This means that their joint distribution factorizes:

$$P(X,Y) = P(X)P(Y)$$

- This factorization is possible because of the chain rule:

$$\begin{aligned} P(X,Y) &= P(X)P(Y|X) \\ &= P(X)P(Y) \end{aligned}$$

i.i.d.

- i.i.d. $=$ independently, identically distributed
- RVs $X_1, ..., X_n$ are i.i.d. iff
  1. They are mutually statistically independent.
  2. All drawn according to the same distribution.

- Note: If $X_1, ..., X_n$ are i.i.d., then

$$\begin{aligned} p(x_1, ..., x_n) &= p_{X_1}(x_1)...p_{X_n}(x_n) \\ &= \prod_{i=1}^{n} p(x_i) \end{aligned}$$

Expectation

- Definition:
$$\mu_x := \mathbb{E}[X] := \int_{\mathcal{X}} xp(x)dx$$
The integral is called the first moment of p.
- Note: Expected value $\neq$ Most likely value.
- For a function f:
$$\mathbb{E}[f(X)] := \int_{\mathcal{X}} f(x)p(x)dx$$

Variance

- Definition:

$$\sigma_X^2 := \mathrm{Var}[X] := \int_{\mathcal{X}} (x - \mu_X)^2 p(x) dx$$

$\rightarrow$ second centralized moment of p.

- Always: $\mathrm{Var}[X] \geq 0$
- Definition: The square root $\sigma_X = \sqrt{\mathrm{Var}[X]}$ is called the standard deviation of X.

Multiple Dimensions

- A vector of random variables

$$\mathbf{X} = (X_1, ..., X_n)^\top$$

  A draw $\mathbf{x} = (x_1 \dots x_n)^\top$ from $\mathbf{X}$ defines a point in n-dimensional space.
- It is treated just like a list of $1D$ RV's.
- The vector components are not necessarily i.i.d
- We can add RV's to produce a new RV

$$Y := c_1 X_1 + c_2 X_2$$

Multidimensional Moment Statistics

- Expectation: Vector of components expectation

$$\mathrm{E}[\mathbf{X}] := (\mathrm{E}[X_1], ..., \mathrm{E}[X_n])^\top$$

- Variance: Generalized to covariance:

$$\begin{array}{rcl} Cov[X,Y] & := & \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y)(x-\mu_X)(y-\mu_Y)dxdy \\ & = & \mathrm{E}_{X,Y}[(x-\mu_X)(y-\mu_Y)] \end{array}$$

- If two RVs have non-zero covariance, we call them correlated
- The covariance is a linear measure statistical dependence

Covariance Behavior

- If X,Y are independent, then $Cov[x, y] = 0$
- Proportional behavior:

$$\begin{aligned} Cov[X, Y] > 0 \quad &\Leftrightarrow \quad X, Y \text{increase together} \\ Cov[X, Y] < 0 \quad &\Leftrightarrow \quad X, Y \text{are anti-proportional} \end{aligned}$$

Covariance Matrix

- For RVs $X_1, ..., X_n$ we use a covariance matrix Σ to describe their mutual covariances:

$$\Sigma_{i,j} := Cov[X_i, X_j] \qquad i, j = 1, ..n$$

- The covariance matrix Σ generalizes the notion of variance to sets of RVs or multiple dimensions.

Covariance Matrix Properties

1. Diagonal entries are RVs variances:

$$\Sigma_{i,j} := Cov[X_i, X_i] = Var[X_i]$$

2. Σ is symmetric:

$$\Sigma_{i,j} = Cov[X_i, X_j] = Cov[X_j, X_i] = \Sigma_{j,i}$$

3. Σ is positive semi-definite

Question: What does a diagonal covariance matrix, Σ mean?

Brain Teaser

Question: Assume you have observed 2D data $\mathbf{X} \in \mathbb{R}^{2\times N}$ (observations as columns). The first row of $\mathbf{X}$ corresponds to the first dimension x_1, the second row corresponds to x_2.

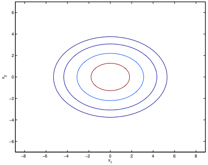
x_1	1.5	4.3	...	0.2
x_2	2.7	-2.1	...	6.0

For each of the 3 covariance matrices $\mathbf{C_X}$, choose the iso-line plot (A-E) corresponding to the covariance matrix.

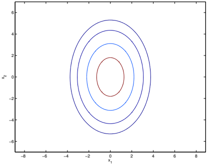
Brain Teaser

1. $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ 2. $\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$ 3. $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

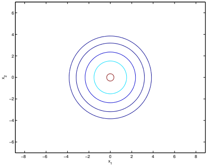
A



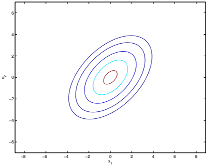
B



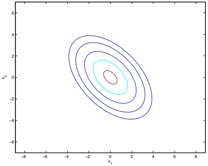
C



D



E



Gaussian Distribution (1D)

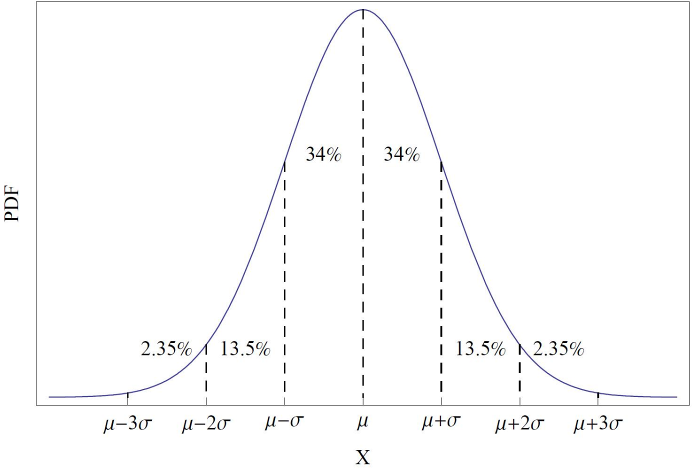
- Sample space $\mathcal{X} = \mathbb{R}$
- Definition: $p(X|\mu, \sigma) := \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(X-\mu)^2}{2\sigma^2})$
- Statistics:

$$\mathrm{E}[X] := \mu, Var[X] := \sigma^2$$

Technically speaking, the Gaussian distribution specifies that the probability density associated with a point x is proportional to the negative exponentiated half-distance to μ scaled by σ^2..

Gaussian Distribution (1D)

Here is a more compelling explanation..



Gaussian Distribution (nD)

- Sample space $\mathcal{X} = \mathbb{R}^n, \mathbf{x} = (x_1, .., x_n)^\top$
- Definition:
  $p(\mathbf{x}|\mu, \Sigma) := \frac{1}{(\sqrt{2\pi})^n |\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu))$

  where Σ is the covariance matrix and $|\Sigma|$ is its determinant

Gaussian Distribution

- Using only correlation/covariance to describe independence means: Higher-order dependencies are neglected.
- This is what the Gaussian does: Parametrized only by location and covariance.
- Describing dependencies in data by covariance is equivalent to approximation of data distribution by a Gaussian model.

Data vs. Distribution

- Important: Be careful to distinguish between distributions (smooth functions in most examples) and data (point clouds).
- Machine learning:
  - Data = input
  - Distribution = model or assumption
- ML methods usually make some general assumptions about distribution, then try to obtain ("infer") the specifics from the data.

Example 1) Modeling step: Assume a Gaussian as model.
2) Inference step: Estimate Gaussian parameters (μ and σ) from data.

Empirical distribution

- We try to regard data sample (imagine some point cloud) as a distribution.
- Problem: We only know whether or not a point is there, not how probable that is.
- Simple solution: Assign same probability to each point.

  Def. Let $S = \{x_1, .., x_n\}$ be a sample of the data, we call

  $$P(x) := \frac{1}{n} \cdot \#\{y \in S | y = x\}$$

  the empirical distribution defined by the data.

Large Sample Theory

Basic question: What can we say about the limiting behavior of a sequence of RVs $X_1, X_2, X_3, \ldots$?
In calculus:

- A sequence of real numbers x_n converges to a limit x if, for every $\epsilon > 0, |x_n - x| < \epsilon$ for all large n
- Trivial example: Suppose $x_n = x$ for all n, then trivially $\lim_{n\to\infty} x_n = x$

In probability theory: for continuous distribution a $\Pr\{X = x_0\} = 0$ thus it's difficult to speak of limits in the same sense as in calculus.

Types of Convergence

Let $X_1, X_2, X_3, \ldots$ be a sequence of RVs and X another RV. Let F_n denote the CDF of X_n and F the CDF of X.

1. X_n converges to X in probability, written $X_n \xrightarrow{P} X$, if for every $\epsilon > 0$

$$\Pr\{|X_n - X| > \epsilon\} \to 0$$

   as $n \to \infty$.

2. X_n converges to X in distribution, written $X_n \rightsquigarrow X$, if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

   at all t for which F is continuous.

Law of Large Numbers (weak statement)

The weak law of large numbers states that the sample average

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges in probability to the expectation

$$\mu = \mathbb{E}(X_i)$$

Question: what conditions are forgotten here for the statement to hold?

Relationships and Transformations

It holds that convergence in

$$\text{probability} \Rightarrow \text{distribution}$$

Some convergence properties are preserved under transformations. Examples:

- If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$.
- The same is not true for convergence in distribution.
- If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$.
- If $X_n \rightsquigarrow X$, then $g(X_n) \rightsquigarrow g(X)$.

Note: Expected Error vs Train Error

Recall the lecture. Training error

$$\hat{R}_D(\mathrm{w}) = \frac{1}{|D|} \sum_{(\mathrm{x},y)\in D} (y - \langle \mathrm{w}, \mathrm{x}\rangle)^2.$$

Note: training error is itself a random variable!
It is exactly the weighted sum from the formulation of the Law of Large Numbers! Its justifies, that when the training set grows, then

$$\hat{R}_D(\mathrm{w}) \xrightarrow{P} R(\mathrm{w}), \quad \text{(weak LLN; in lecture — strong LLN).}$$

Be careful: in most cases the expected value of the training error underestimates the expected error of the training set:

$$\mathbb{E}\, \hat{R}_D^{\text{train}}(\mathrm{w}) \leq \mathbb{E}\, \hat{R}_D^{\text{test}}(\mathrm{w}).$$

Reason: training set is used for training!
Can be proven rigorously in case of regression (see in later tutorial).