Introduction to Pattern Recognition: Lectures 2 and 3: Probability and Statistics

Jussi Tohka

jussi.tohka@tut.fi

Institute of Signal Processing

Introductory example

From http://davidmlane.com/hyperstat/probability.html
What is the probability that a card drawn at random from a deck of cards will be an ace?

- Since of the 52 cards in the deck, 4 are aces, the probability is 4/52.
- In general, the probability of an event is the number of favorable outcomes divided by the total number of possible outcomes. (This assumes the outcomes are all equally likely.)
- In this case there are four favorable outcomes: (1) the ace of spades, (2) the ace of hearts, (3) the ace of diamonds, and (4) the ace of clubs. Since each of the 52 cards in the deck represents a possible outcome, there are 52 possible outcomes.

Introductory example 2

The same principle can be applied to the problem of determining the probability of obtaining different totals from a pair of dice. As shown below, there are 36 possible outcomes when a pair of dice is thrown.

Die 1	Die 2	Total
1	1	2
1	2	3
1	3	4
1	4	5
1	5	6
1	6	7
2	1	3
2	2	4
2	3	5
2	4	6
2	5	7
2	6	8
3	1	4
3	2	5
3	3	6
3	4	7
3	5	8
3	6	9

Die 1	Die 2	Total
4	1	5
4	2	6
4	3	7
4	4	8
4	5	9
4	6	10
5	1	6
5	2	7
5	3	8
5	4	9
5	5	10
5	6	11
6	1	7
6	2	8
6	3	9
6	4	10
6	5	11
6	6	12

Introductory example 2 (cont.)

- To calculate the probability that the sum of the two dice will equal 5, calculate the number of outcomes that sum to 5 and divide by the total number of outcomes (36).
- Since four of the outcomes have a total of 5 (1,4; 2,3; 3,2; 4,1), the probability of the two dice adding up to 5 is 4/36 = 1/9 .
- In like manner, the probability of obtaining a sum of 12 is computed by dividing the number of favorable outcomes (there is only one) by the total number of outcomes (36). The probability is therefore 1/36 .

Probability

- Probability theory plays a central role in pattern classification. Indeed, rarely two patterns from the same category (or class) appear to be exactly same. Instead, usually there is some variation within the patterns drawn from a certain category. To model this variation, we must introduce *probabilistic model* for the category.
- However, to obtain a more useful definition of probability than the simple one provided above, we must introduce some (easy) mathematics that will come useful later on.

Probability

Note: Some definitions may appear unnaturally abstract and far-fetched, but the intention of these two lectures is not to scare you nor provide you with the ability to prove theorems and lemmas, but deepen the understanding of the concept of probability. That is, the definitions themselves are not important, the important thing is what they mean. Hopefully, this insight will enable the understanding why we do what we do later on.

Probability: Basic definitions

- **The sample space**: The set S is a sample space for an experiment (for us a measurement), if every physical outcome of the experiment refers to a unique element of S.
- **The event**: A subset of the sample space. If two events do not intersect, i.e their intersection is empty, they are called *mutually exclusive*. More generally, events $E_1, E_2 \dots E_n$ are mutually exclusive if $E_i \cap E_j = \emptyset$ for any $i \neq j$.

Probability: Basic definitions

- **The probability space**: Given a sample space S, a *probability measure* P on S is a real-valued function defined on the events of S such that
  1. For any event $E \subseteq S$, $0 \leq P(E) \leq 1$.
  2. P(S) = 1.
  3. If events $E_1, E_2, \dots, E_i, \dots$ are mutually exclusive

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

- Venn diagrams

Probability: Examples

- Lets go back to our preliminary examples and analyze them in the light of the definitions in the previous slide. In the first example, we were asked *What is the probability that a card drawn at random from a deck of cards will be an ace?*
- The (most natural) sample space in this case consists of the labels of cards, e.g. 3 of spades, 7 of hearts, ace of clubs. The event which we are interested is then {ace of spades, ace of clubs, ace of hearts, ace of diamonds }.

Probability: Examples

- The (most natural) sample space in this case consists of the labels of cards, e.g. 3 of spades, 7 of hearts, ace of clubs. The event which we are interested is then {ace of spades, ace of clubs, ace of hearts, ace of diamonds }.
- The probability measure is defined by $P(\text{any} \quad \text{card}) = 1/52$. Symbol as is used for the ace of spades, ac for ace of clubs etc. Now, the asked probability is (based on the property 3)

$$\begin{aligned} P(\text{ace}) &= P(\text{as} \cup \text{ac} \cup \text{ah} \cup \text{ad}) \\ &= P(\text{as}) + P(\text{ac}) + P(\text{ah}) + P(\text{ad}) = 4/52. \end{aligned}$$

Properties of probability spaces

Probability spaces have some nice and useful properties, which can be proved from the three axioms only. To present (and prove) these properties one must rely mathematical abstraction of the probability spaces offered by their definition. However, as we are more interested in their interpretation, a little dictionary is needed: In following E and F are events and E^c denotes complement of E.

set theory	probability
E^c	E does not happen
$E \cup F$	E or F happens
$E \cap F$	E and F both happen

Properties of probability spaces

1. $P(E^c) = 1 - P(E)$
2. $P(\emptyset) = 0$
3. If E is a sub-event of F (i.e. $E \subseteq F$), $P(F - E) = P(F) - P(E)$
4. $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

set theory	probability
E^c	E does not happen
$E \cup F$	E or F happens
$E \cap F$	E and F both happen

Conditional Probability

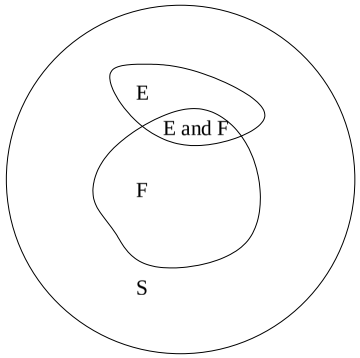
Sometimes, in fact quite often, we are interested in probability of the event E provided that the event F has already happened. This probability, called conditional probability relative to $F \subseteq S$ is denoted by $P(E|F)$ and defined by

$$P(E|F) = \frac{P(E \cap F)}{P(F)},$$

for any event $E \subseteq S$ and provided that $P(F) > 0$. $P(E|F)$ is read *probability of* E *given* F. (Note that $P(\cdot|F)$ for a fixed event F with a nonzero probability is a legitimate probability measure on S.)

Conditional Probability

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$



Bayes Theorem

Now we are in position that we can introduce some very important properties of the conditional probability.

- A *partition of a sample space* S is a collection of mutually exclusive events $F_1, \ldots, F_n$ such that $S = F_1 \cup \cdots \cup F_n$. For any partition $F_1, \ldots, F_n$ of a sample space S and any event $E \subseteq S$ it holds that

$$P(E) = \sum_{i=1}^{n} P(E \cap F_i).$$

- **The multiplication property** states that for events E, F, if $P(F) > 0$,

$$P(E \cap F) = P(F)P(E|F).$$

Bayes Theorem

- **The Bayes theorem**. Suppose that the events $F_1, F_2, \ldots, F_n$ is a partition of the sample space and E is an event with $P(E) > 0$. Then for any k,

$$P(F_k|E) = \frac{P(F_k)P(E|F_k)}{P(E)} = \frac{P(F_k)P(E|F_k)}{\sum_{i=1}^{n} P(F_i)P(E|F_i)},$$

particularly,

$$P(F|E) = \frac{P(F)P(E|F)}{P(E)},$$

for any event F with $P(F) > 0$

- The Bayes' theorem is often referred as the Bayes rule, and to get everyone confused, it can be written in several different (but essentially the same) ways.

Independence

- Events E and F in S are *independent* if

$$P(E \cap F) = P(E)P(F).$$

  If E and F are not independent they are *dependent*.
- **Theorem.** If E and F are independent, then $P(E|F) = P(E)$ and also E and F^c are independent, and so are E^c and F^c, E^c and F.
- Independence of events means that occurrence of E does not affect the likelihood of occurrence of F. Or more generally, the occurrence of one or more events does not affect the likelihood of the occurrence of the remaining ones. For instance, events that follow from repeating the same experiment several times are usually independent.

Independence

Assume that we have more than two events, say E, F and G. They are pairwise independent if E and F are independent and F and G are independent and E and G are independent. However, **it does NOT follow from the pairwise independence that** $P(E \cap F \cap G) = P(E)P(F)P(G)$. The definition of independence of several events is somewhat lengthy. However, if we say that events $E_1, \ldots, E_n$ are independent then $P(E_1 \cap \cdots \cap E_n) = P(E_1) \cdots P(E_n)$.

Random variables

- *Random variables* characterize the concept of random measurement in probability theory.
- **A random variable (RV) is a real-valued function X defined on a sample space S, i.e. $X : S \to \mathbb{R}$.** The set of values $\{X(x) : x \in S\}$ taken on by X is called the *co-domain*.
- Recall our first example, where we wanted to compute the probability of drawing an ace from a deck of cards. What is a relevant random variable in this case? The sample space consisted of the labels of the cards. The random variable in this case is X : labels of cards $\to \{0, 1\}$, where $X(\text{ace}) = 1$ and $X(\text{other} \quad \text{cards}) = 0$.

Random variables

- In the second example, a relevant RV would be the mapping from the sample space consisting of pairs where each element is an integer between 1 and 6 to the sum of these elements.

RVs and probability

- Given a random variable X a basic problem is to give meaning to probability statements of the form $P(A_{a,b}) = p$, where (event) $A_{a,b} = \{x : a < X(x) < b\}$.
- Note that notation $P(a < X < b)$ is used more than often to denote $P(A_{a,b})$.
- More generally we are interested in statements like

$$P(X \in B) = P(\{x : X(x) \in B \subseteq \mathbb{R}\}).$$

RVs and probability

- A concept related to RVs is the **probability distribution function** (or cumulative distribution function (cdf)). For a RV X, its probability distribution function F_X is defined for all real x as

$$F_X(x) = P(X \leq x).$$

- Often we are little bit sloppy in notation and denote F_X simple by F.
- Properties of the cdfs include:
  1. cdfs are increasing, i.e. if $x \leq y$, then $F_X(x) \leq F_X(y)$.
  2. $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$.

Probability densities:Discrete case

- A RV X is said to be *discrete* if its co-domain is denumerable (i.e. points in the co-domain can be written as a list, finite or infinite).
- Its **probability density function (pdf)** f_X is then defined as

$$f_X(x) = \begin{matrix} P(X = x) & \text{if } x \text{ is in co-domain of } X \\ 0 & \text{otherwise} \end{matrix}.$$

  And again, we often omit the subscript X when denoting the density.
- Write co-domain of X as $\{x_1, x_2, \ldots, \}$. Then

$$P(X \leq x) = F_X(x) = \sum_{x_k \leq x} f_X(x_k).$$

Probability densities:Discrete case

- Lets once again consider the deck of cards and the probability of drawing an ace. The RV in this case was the map from the labels of the cards to the set $\{0, 1\}$ with value of the RV equal to 1 in the case of a ace and 0 otherwise.
- Then $F_X(x) = 0$, when $x < 0$, $F_X(x) = 48/52$, when $0 \leq x < 1$, and $F_X(x) = 1$ otherwise.
- Density $f_X(0) = 48/52$ and $f_X(1) = 4/52$ and $f_X(x) = 0$ otherwise.

Probability densities: Continuous case

- Another important class of RVs is comprised of those which are continuous. These take values on intervals of real line.
- The probability density function f_X of a continuous RV X is such that

$$P(X \in B) = \int_B f_X(x)dx.$$

- Actually, continuous RVs are defined to be those which have pdf satisfying the above equation. (So we do not have to worry about integrability.)

Probability densities: Continuous case

- **Theorem**. A real-valued function f is the probability density for a continuous RV if and only if
  1. $f(x) \geq 0$ for all x.
  2. $\int_{-\infty}^{\infty} f(x)dx = 1$.
- For a continuous RV X it holds that the derivative

$$F'_X(x) = f_X(x),$$

  if F_X is continuous at x.

Distributions, densities and probability

- With probability densities at our disposal, we often view whole measurement process through them.
  - RVs and their densities are often identified and sample spaces are left unspecified.
  - Random variables are often dropped from our notation, e.g. we write $P(X \in B)$ as $P(B)$ as if B was an event (which it, directly, is not).

Distributions, densities and probability

- One might then ask 'Why all this trouble?'
  - Rigorous formulation of concepts gives a firm ground to build all further mathematics.
  - True, but that is not what we are after in this course.
  - Instead in pattern recognition one makes measurements about the object in order to recognize the object. The recognition process with computers requires mathematical abstraction in order to develop efficient methods and algorithms. The input and output of this process, however, are not mathematical abstractions. Thus we need machinery to build probabilistic models from basics and then interpret the results, and that is what we are after in developing a little bit of probability theory

Random vectors

- It is common and useful to consider two or more RVs jointly. In such a circumstance, it is not merely a question of this or that random variable taking a particular value, but rather the likelihood of a number of random variables taking a number of values. Therefore, we cannot consider random variables in isolation of each other; we must take also their joint characteristics into account.
- If we have n somehow related random variables $X_1, X_2, \ldots, X_n$ resulting from some measurement process, we consider a *random vector* $X = (X_1, X_2, \ldots, X_n)$. Random vectors differ from univariate RVs only in the respect that they take on values from $\mathbb{R}^n$ instead of $\mathbb{R}$.
- From now on a random variable can mean a random vector taking on values from $\mathbb{R}^n$.

Probability densities: multivariate case

Random vectors have also associated probability densities.

- In the discrete case, the **joint probability density function** of RVs $X_1, \ldots, X_n$ is defined by

$$f_{(X_1,\ldots,X_n)}(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n),$$

  for every $(x_1, \ldots, x_n)$ in the co-domain of $X = (X_1, X_2, \ldots, X_n)$.
- The probability density function of a random vector $X = (X_1, X_2, \ldots, X_n)$ is written as

$$f_X(\mathbf{x}) = P(X = \mathbf{x}).$$

- Both definitions mean the same thing!

Probability densities: multivariate case

In the continuous case the joint pdf $f_{X_1,\ldots,X_n} = f_X$ of $X = (X_1, \ldots, X_n)$ satisfies

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = \int_{B_1} \cdots \int_{B_n} f_{(X_1,\ldots,X_n)}(x_1, x_2, \ldots, x_n) dx_1 \cdots dx_n,$$

or more nicely

$$P(X \in B) = \int_B f_X(\mathbf{x}) d\mathbf{x}.$$

Distribution functions: Multivariate case

- As one could guess from the notation used in the previous slide, most things in the multivariate case work principally in the same way as in the univariate case.
- For example, the joint distribution function (cdf) for a vector valued and continuous RV X,

$$F_X(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f_X(\mathbf{x}) d\mathbf{x},$$

- The density is obtained from cdf by differentiating.
- **Do not be afraid of the multiple integrals, you do not have to compute them!**

Joint distributions of random vectors

- Similarly as univariate random variables, vector-valued RVs may have a joint distribution.
- Let X and Y be vector valued RVs, then their joint density function satisfies

$$P(X \in A, Y \in B) = \int_A \int_B f_{X,Y}(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}.$$

- Also, it is possible to recover densities of X and Y when given the joint density. These are called **marginal densities**. When X and Y are jointly continuous with density $f_{X,Y}$, the marginal density of X is given by

$$f_X(\mathbf{x}) = \int_{-\infty}^{\infty} f_{X,Y}(\mathbf{x}, \mathbf{y}) d\mathbf{y}.$$

Independence of RVs

- Random variables (real or vector valued) $X_1, \ldots, X_n$ possessing joint density f are said to be independent if

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = f_{X_1}(\mathbf{x}_1) \cdots f_{X_n}(\mathbf{x}_n).$$

- **Theorem**. RVs $X_1, \ldots, X_n$ are independent if and only if

$$F(\mathbf{x}_1, \ldots, \mathbf{x}_n) = F_{X_1}(\mathbf{x}_1) \cdots F_{X_n}(\mathbf{x}_n),$$

  with obvious notation.
- Also, for independent RVs it holds that

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n).$$

Conditional distributions

- We have already dealt with conditional probability, the probability of an event provide that another event has happened.
- An analogous definition can be made also for probability densities. Question asked this time around is 'What can be said about Y, given we know outcome of X?
- **Conditional density** of Y given $X = \mathbf{x}$

$$f(\mathbf{y}|\mathbf{x}) = \frac{f_{X,Y}(\mathbf{x}, \mathbf{y})}{f_X(\mathbf{x})}.$$

- The related random variable is called the **conditional random variable of Y given** x and denoted by $Y|\mathbf{x}$.

Bayes theorem

- The multiplication property and the Bayes theorem also have their counterparts when dealing with densities. (We shall omit the subscripts when denoting densities.)
- **The multiplication property**: $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{y}) f(\mathbf{x}|\mathbf{y})$.
- **The Bayes theorem**: $f(\mathbf{y}|\mathbf{x}) = \frac{f(\mathbf{x}|\mathbf{y}) f(\mathbf{y})}{\int_{-\infty}^{\infty} f(\mathbf{x}|\mathbf{y}) f(\mathbf{y}) d\mathbf{y}}$.

Expectation and variance

- Two important topics are still untouched. These are the expected value and variance of a random variable.
- Imagine that the same experiment is run infinitely many times. Then the expected value provides us with information what is the value resulting from experiment on average. And, the variance provides us with a measure of how much these experiment results differ on average.

Expectation and variance

- For a RV X with density f, **the expected value** of X is defined as

$$E[X] = \int_{-\infty}^{\infty} \mathbf{x} f(\mathbf{x}) d\mathbf{x},$$

- **The variance** of X is

$$Var[X] = E[(\mathbf{x} - E[X])(\mathbf{x} - E[X])^t].$$

- Note that in the case of vector valued RV, the expected value is a vector and the variance is a matrix.