Some notes on Machine Learning

January 17, 2019

1 Conventions

Let X_{ij} be a $n \times p$ Matrix with n samples and p features.

- $X_{.j} = \sum_{i=1}^{n} X_{ij}$
- $X_{i.} = \sum_{j=1}^{p} X_{ij}$

2 Some Basics

Definition 1 (Metric)**.** *Let X be a set. A metric is a map $d : X \times X \to \mathbb{R}_{\geq 0}$ $:\Leftrightarrow$ $\forall x, y \in X$ the following properties are fulfilled:*

1. $d(x, y) = 0 \Leftrightarrow x = y$
2. $d(x, y) = d(y, x)$ *(symmetry)*
3. $d(x, y) \leq d(x, z) + d(z, y)$ *(triangle inequality)*

Definition 2 (Norm)**.** *Let V be a vector space. A norm is a map $\|\cdot\| : V \to \mathbb{R}_{\geq 0}$ $:\Leftrightarrow$ $\forall x, y \in V$ and $\alpha \in \mathbb{R}$ the following properties are fulfilled:*

1. $\|x\| \Leftrightarrow x = 0$
2. $\|\alpha x\| = |\alpha| \|x\|$
3. $\|x + y\| \leq |x| + \|y\|$ *(triangle inequality)*

A pair $(V, \|\cdot\|)$ is called normed vector space. The notions of norm and metric are closely related. Every norm induces a metric.

Theorem 1. *Let $(V, \|\cdot\|)$ be a normed vector space. Then*

$$d(x, y) := \|x - y\| \tag{1}$$

is called the metric induced from the norm.

Note that not every metric is induced from a norm – for example, the discrete metric in Eq. 16.

One important norm is the so-called p-norm

Definition 3 (p-norm)**.** *Let $x \in \mathbb{R}^n$ and $1 \leq p < \infty$*

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \tag{2}$$

It can be shown that for $1 \leq p < \infty$ the p-norm is indeed a norm, i.e., obeys definition 2. Furthermore, p-norms are equvalent norms. However, for $0 < p < 1$ it is only a quasi-norm because the triangle inequality is no longer fulfilled.[1]

Definition 4 (Scalar product)**.** *Let V be a vector space. A scalar product is a map $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R} :\Leftrightarrow \ \forall x, y, z \in V$ and $\alpha, \beta \in \mathbb{R}$ the following properties are fulfilled:*

1. $\langle x, x \rangle \geq 0$
2. $\langle x, x \rangle = 0 \Leftrightarrow x = 0$
3. $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$
4. $\langle x, y \rangle = \langle \bar{y, x} \rangle$

A pair $(V, \langle \cdot, \cdot \rangle)$ is called inner product space. Inner product spaces are normed spaces are again related.

Theorem 2. *Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space. Then*

$$\|x\| = \sqrt{\langle x, x \rangle} \tag{3}$$

is the norm induced from the inner product.

3 Dissimilarities and Similarities

Closely related to the concept of metric is the concept of dissimilarity. For details see Ref. [8], which contains also links between dissimilarity, similarity, and metric.

Definition 5 (Dissimilarity function)**.** *Let X be a set. A dissimilarity function is a map $d : X \times X \to \mathbb{R}_{\geq 0} :\Leftrightarrow \ \forall x, y \in X$ the following properties are fulfilled [8]:*

1. $d(x, y) = 0 \Leftrightarrow x = y$

4 Bayes' Rule

Theorem 3 (Bayes)**.**

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{4}$$

Proof. Use $P(x, y) = P(x|y)P(y) = P(y|x)P(x)$ □

[1] However, for a quasi-norm $\exists \lambda \in \mathbb{R}$ such that $\forall x, y \in X$ we have $\|x + y\| \leq \lambda \left(\|x\| + \|y\| \right)$

5 k-prototype algorithm

This method is suitable for clustering data sets, which contain both, numerical and categorical variables. The goal of cluster analysis is to partition the observations into groups ("clusters") so that the pairwise dissimilarities between those assigned to the same cluster tend to be smaller than those in different clusters.

Consider a feature space that consists of a (vector space) of numerical features V and categorical features C,

$$F = V \times \mathcal{C} \tag{5}$$
$$V = \mathbb{R}^n \tag{6}$$
$$C = \mathcal{C}_1 \times \mathcal{C}_2 \times \ldots \times \mathcal{C}_d \tag{7}$$

In order to obtain a neighborhood-based clustering method for such a mixed type of feature space we first consider numerical and categorical variables individually. The former will lead to the k-means algorithm and the latter to the k-mode algorithm. Combining both algorithms will lead to the k-prototype algorithm.

5.1 Formal setup

The most popular clustering algorithms directly assign each observation to a group or cluster without regard to a probability model describing the data. Each observation $x_i \in F$ is uniquely labeled by an integer $i \in \{1, \ldots, N\}$. A prespecified number of clusters $K < N$ is postulated, and each cluster C_k one is labeled by an integer $k \in \{1, \ldots, K\}$. For each x_i we introduce an indicator function with respect to membership of cluster C_k,

$$r_{ik} = \begin{cases} 1 & x_i \in C_k \\ 0 & x_i \notin C_k \end{cases} \tag{8}$$

This is also known as 1-of-k encoding, because a given data point may not belong to two distinct clusters.

In order to introduce a distance from a cluster to a datapoint, a prototype (sometimes also called cluster centroid or cluster center) q_k for each cluster is introduced. The distance between a data point x_i and a cluster is then defined as the distance between data point and cluster prototype $d(x_i, q_k)$, where d is sometimes also called distortion measure. This allows to define an objective function,

$$J = \sum_{k=1}^{K} \sum_{i=1}^{N} r_{ik} d(x_i, q_k), \tag{9}$$

which contains sums over all N datapoints and K clusters. The clustering algorithm is then just minimizing the cost function!

For a fixed metric d, the cost function depends then on the centroids and the indicator function, $J = J(q, r)$. Eq. (9) may be minimized iteratively

1. minimize r keeping q fixed

2. minize q keeping r fixed

The minimization of r for a fixed metric q may be obtained by

$$r_{ik} = \begin{cases} 1 & \text{if } k = \text{argmin}_l \, d(x_i, q_l) \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Thus, each datapoint is assigned to the closes centroid with respect to the metric d.

The following algorithms are just different incarnation of minimizing the cost function J for different dissimilarity measures.

5.2 K-means algorithm

Consider only numerical features, $F = \mathbb{R}^n$. Therfore, we have $x_i, q_k \in \mathbb{R}^n$ The metric for the cost function may be chosen by the squared euclidean distance (2-norm),

$$d(x_i, q_l) = \|x_i - q_l\|_2^2 \tag{11}$$

The cost function (9) then assumes the form

$$J = \sum_{k=1}^{K} \sum_{i=1}^{N} r_{ik} \|x_i - q_k\|_2^2. \tag{12}$$

The iterative solution from above is now of the form,

1. minimize r keeping q fixed:

$$r_{ik} = \begin{cases} 1 & \text{if } k = \text{argmin}_l \|x_i - q_l\|_2^2 \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

2. minize q keeping r fixed:

$$q_k = \frac{\sum_{i=1}^{N} r_{ik} x_i}{\sum_{i=1}^{N} r_{ik}} = \frac{\sum_{i=1}^{N} r_{ik} x_i}{n_k} \tag{14}$$

with n_k being the number of members in in C_k.

Proof. Consider q_{kj}, the j-th component of q_k. With this notation x_{ij} is the j-th component of datapoint i. Then,

$$\frac{\partial J}{\partial q_{kj}} = -\sum_{i=1}^{N} 2 r_{ik} (x_{ij} - q_{kj}) \stackrel{!}{=} 0 \Rightarrow \sum_{i=1}^{N} r_{ik} x_{ij} = \sum_{i=1}^{N} r_{ik} q_{kj}$$

□

Note that that Eq. (14) is just the mean of all datapoints in C_k.

5.2.1 K-means and categorical features using Gower

TODO!!!
Within the k-means approach only numerical features are tractable. However, also treat categorical features using the Gower metric

5.3 K-mode algorithm

Now consider the case of M categorical features only, $F = \mathcal{C} = \mathcal{C}_1 \times \mathcal{C}_2 \times \cdots \times \mathcal{C}_d$. Consider two data points $x_n, x_m \in \mathcal{C}$, and thus, $x_{ni}, y_{mi} \in \mathcal{C}_i$. One dissimilarity measure on categorical variables is the so called simple matching [4, 3] (sometimes also called discrete metric),

$$d_1(x_n, x_m) = \sum_{i=1}^{d} \delta(x_{ni}, y_{mi}) \tag{15}$$

$$\delta(x_{ni}, y_{mi}) = \begin{cases} 0 & \text{if } x_{ni} = y_{mi} \\ 1 & \text{if } x_{ni} \neq y_{mi}. \end{cases} \tag{16}$$

So, $\delta(x_{ni}, y_{mi})$ measures the dissimilarity of the categorical feature i for datapoints x_n and x_m and $d_1(x_n, x_m)$ is the sum over all f categorical features. The cost function (9) then assumes the form

$$J = \sum_{k=1}^{K} \sum_{i=1}^{N} r_{ik} d_1(x_i, q_k), \tag{17}$$

where q_k is the centroid of cluster C_k. The double sum runs over all K centroids and N data points. Again, we neet to evaluate the two optimization steps. To this end, we need a piece of theory [4].

- define frequency here

Definition 6. *Let $x = \{x_1, x_2 \ldots, x_N\}$ be a sample, and let $s_i \in \mathcal{C}$.*

- *The **frequency** of s_i given sample x is defined as*

Theorem 4. *q is a minimizer of the function*

$$\mathcal{C} \ni q' \mapsto \sum_{i=1}^{N} d_1(x_i, q') \tag{18}$$

$\Leftrightarrow$ q is a mode of x, i.e., $f_r(q_i|x) \geq f_r(s_i|x),\ \forall s_i \in \mathcal{C}_i,\ \forall i \in 1 \ldots d$.

Proof.

$$\sum_{i=1}^{N} d_1(x_i, q') = \sum_{j=1}^{d} \sum_{i=1}^{N} \delta(x_{ij}, q'_j) = \sum_{j=1}^{d} n(1 - f(q'_j|x)).$$

Because $n(1 - f(q'_j|x)) \geq 0$ for $1 \leq j \leq d$, Eq. (18) mininimizes $\Leftrightarrow$ each $n(1 - f(q'_j|x))$ minimizes individually. Thus, $f(q'_j)$ maximizes and therefore q is a mode. □

So, q is determined by assigning the most frequent attributes - but this is just the mode. However, note that the mode is not unique.

Method	optimal value
Gap statistics	Elbow
Silhouette index	Max
Dunn index	Max
Calinski-Harabasz index	Max
Davies-Bouldin index	Min

Table 1: Compilation of internal clustering validation measures [6].

5.4 K-prototype Algorithm for p-norms

We have seen that the k-means algorithm hinges on the Euclidean norm. Because then the prototype of a cluster is the mean of the cluster. Here we study the k-prototype algorithm for p-norms from Eq. (2).

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

However, we extend the values for p, $0 < p < \infty$. So, strictly speaking this is not a norm (in the case for $p < 1$) but it is still a dissimilarity function. We consider the class of cost functions

$$J = \sum_{k=1}^{K} \sum_{i=1}^{N} r_{ik} \|x_i - q_k\|_p^m. \tag{19}$$

So, for $m = p = 2$ this reduces to the K-means. Before we go on, lets collect our basic assumptions:

- K-means works with dissimilarities as well (then we are allowed to consider also $0 < p < 1$ values)
- K means also works with powers other than $m = 2$.

These should be carefully checked at some point. For this generalization the K-prototype algorithm than assumes the form:

1. minimize r keeping q fixed:

$$r_{ik} = \begin{cases} 1 & \text{if } k = \text{argmin}_l \|x_i - q_l\|_p^m \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

2. minize q keeping r fixed: q_k needs to be determined numerically. No closed analytical solution may be found in general, because the absolute value part in the p-dissimilarity part (only for p even and > 0).

5.5 Determing the number of clusters

There are several methods to obtain the optimal number of clusters [6], which are partially compiled in Table 1.

Gap statistics aka ellbow method: In the above approaches K, the number of clusters enters as a parameter. In principle determining K depends on the goal. For example, if we want to assign Costumers to p given print campaigns then it is natural to choose $p = K$. On the other hand, if we want to study the inherent clustering structure of a dataset we need a guideline for choosing K.

One way is to consider the (minimised) cost function $J(K)$ as function of the number of Clusters. K is then selected where the cost function shows kinks (if there exist any). The rationale behind this idea is that one assumes that the data has some inherent clustering structure with K^* clusters. For $K < K^*$ additional clusters would decrease the cost function (which is always monotonically decreasing with cluster size) quite significantly because we approach the inherent cluster structures: $J_K << J_{K+1}$, for $K < K^*$. For $K \geq K^*$ the decrease of the cost function should be much smaller, $J_K < J_{K+1}$ because the additional sub-clusters should be close within an inherent cluster. Thus, at K^* we are expecting a sharp peak in the successive differences of the cost function,

$$\{J_K - J_{K+1} | K < K^*\} << \{J_K - J_{K+1} | K \geq K^*\} \tag{21}$$

Gap statistics further refines this idea and builds on $\log(J(K))$ [7]. It is also applicable for hierarchical clustering.

Silhouette Coefficient:

5.6 Missing values

In real life data often values are missing. Assume that we have an observation $x_i \in F$ with a missing value for attribute x_{ij}. Some standard approaches in dealing with this problem are: [2]

- Omit the contribution of Attribute j to the dissimilarity $d(x_i, x_i')$ between the two objects x_i, x_i'. This is the most common approach. However, it may fail if both observations have no common values at all. Then the two observations could be deleted from the analysis.
- Missing values could be replaced by the mean or median of each attribute over the non-missing data.
- For categorical variables 'missing' could be interpreted as just another category. This is reasonable if one can consider two objects as similar if they have both missing values.

6 Hyptothesis testing

6.1 Analysis of variance: ANOVA

The t-test allows for a comparison of the means of two independent random samples (categories). The basic idea behind the t-test is to decide via the sample means if the two samples are significantly different.

However if the number of samples $I > 2$, then the t-test is doomed to fail. The ANOVA allows for $I > 2$.

Definition 7. *A **factor** is an independent variable that is contained in all I categories. The I individual categories are called **Factor levels**.*

Null-hypothesis. The null hypothesis for ANOVA is that all samples have the same mean,

$$H_0 : \ \mu_1 = \mu_2 = \ldots \mu_I \tag{22}$$

Assumptions.

- The random samples are independent
- All random samples have the same variance (variance homogeneity): $\sigma_i = \sigma \ \forall i \in 1 \ldots I$
- The ith sample is a Normal distribution $\mathcal{N}(\mu_i, \sigma^2) \ \forall i \in 1 \ldots I$

What happens if my data fail these assumptions? Firstly, don't panic! The first two of these assumptions are easily fixable, even if the last assumption is not. Lets go through the options as above:

- The one-way ANOVA is considered a robust test against the normality assumption. This means that it tolerates violations to its normality assumption rather well. As regards the normality of group data, the one-way ANOVA can tolerate data that is non-normal (skewed or kurtotic distributions) with only a small effect on the Type I error rate. However, platykurtosis can have a profound effect when your group sizes are small. This leaves you with two options: (1) transform your data using various algorithms so that the shape of your distributions become normally distributed or (2) choose the non-parametric Kruskal-Wallis H Test which does not require the assumption of normality.
- There are two tests that you can run that are applicable when the assumption of homogeneity of variances has been violated: (1) Welch or (2) Brown and Forsythe test. Alternatively, you could run a Kruskal-Wallis H Test. For most situations it has been shown that the Welch test is best. Both the Welch and Brown and Forsythe tests are available in SPSS Statistics (see our One-way ANOVA using SPSS Statistics guide).
- A lack of independence of cases has been stated as the most important assumptions to fail. Often, there is little you can do that offers a good solution to this problem.

6.2 Kruskal Wallis

The Kruskal Wallis (KW) test [5] is a parameter free test. It tests whether independent samples come from the same underlying population.

- The KW test only answers the question if a sets of samples come from a joint population.
- Post-hoc tests are necessary to compare the sample sets pairwise and determine which of the samples are different. Such multiple hypothesis tests require a careful treatment of the familywise error rate via, e.g., the Šidák or Bonferroni method.

7 Moment generating function

Definition 8 (Moment generating function)**.** *The moment generating function* $M_X : \mathbb{R} \to \mathbb{R}$ *of a real valued random variable* X *over some probability space* (Ω, Σ, P) *is defined by*

$$M_X(t) = E_X(e^{tX}) = \int_\Omega e^{tx} dP, \tag{23}$$

where $E_X(\cdot)$ *is the expectation value with respect to random variable* X*.*

Apparently this is a function of t and can be rewritten using the PDF $f_X(x)$,

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx, \tag{24}$$

Generalisation to more than one dimension. The generalisation of definition 8 to l dimensions is straightforward by replacing tX by the scalarproduct. Let $X = (X_1, X_2, \ldots, X_l)$ and $M_X(t) = M_X(t_1, t_2, \ldots t_l)$

$$M_X(t) := E_X\left(e^{\langle X,t\rangle}\right) \tag{25}$$

Theorem 5 (Moment generation)**.** *The* k*-th moment* $m_k := E(X^k)$ *is generated by the* k*-fold derivative of* $M_X(t)$

$$m_k = \frac{d^k}{dt^k} M_X(t)\bigg|_{t=0} \tag{26}$$

Proof. We express the exponential in a Taylor series,

$$M_X(t) = \int_{-\infty}^{\infty} \left[1 + tx + \frac{t^2}{2!}x^2 + \ldots\right] f(x) dx \tag{27}$$

□

Remark Most of the time we are not interested in moments but rather in central moments (i.e., moments with respect to the mean). $\ln M_X(t)$ would generate these.

Theorem 6 (Independet variables)**.** *Let* $X_i, i = 1 \ldots n$ *be* n *independent random variables. Then* $M_Y(t)$ *for the sum* $Y = \sum X_i$ *factorizes,*

$$M_Y(t) = M_{X_1}(t) \ldots M_{X_n}(t) \tag{28}$$

Proof.

$$M_Y(t) = E(e^{tY}) = E(e^{tX_1} \ldots e^{tX_n}) = E(e^{tX_1}) \ldots E(e^{tX_n}) = M_{X_1}(t) \ldots M_{X_n}(t)$$

where we have used that for independent variables the expectation value factorizes. □

Now, we come to an interesting question: how much do we know about the underlying PDF once the MGF is computed? The answer is all of it! This is really valuable because often it's much easier to deal with MGFs than with the PDF and in addition we get the moments for almost free.

Theorem 7 (Uniqueness)**.** *If the moment generating function exists in a neighbourhood of 0, then the PDF is uniquely determined.*

More precisely let X and Y be random variables and $\exists\ \varepsilon > 0$ such MGFs $M_X(t), M_Y(t) < \infty$ for $t \in (-\varepsilon, \varepsilon)$. Then $f_X = f_Y \Leftrightarrow M_X(t) = M_Y(t)\ \forall t \in (-\varepsilon, \varepsilon)$. A proof can be found in Ref.[1].

8 Characteristic function

Definition 9 (Characteristic function)**.** *The characteristic function* $\Phi_X : \mathbb{R} \to \mathbb{C}$ *of a real valued random variable* X *over some probability space* (Ω, Σ, P) *is defined by*

$$\Phi_X(t) = E_X(e^{itX}) = \int_\Omega e^{itx} dP, \tag{29}$$

where $E_X(\cdot)$ *is the expectation value with respect to random variable* X*.*

So, it is essentially the Fourier transform. Alternatively, these expressions may be obtained via the CDF F_X, the PDF f_X, or in the discrete case:

- via the CDF: $\Phi_X(t) = \int_{-\infty}^{\infty} e^{itx} dF_X(x)$
- via the PDF: $\Phi_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(x) dx$
- discrete case: $\Phi_X(t) = \sum_k e^{itx_k} P(X = x_k)$

Example standard-normal-distribution. Let $X \sim \mathcal{N}(0, 1)$. Then the characteristic function is given by:

$$\Phi_X(t) = e^{-\frac{t^2}{2}}, \tag{30}$$

Proof. Instead of performing the complex integration of Definition 9 we rather use a beautiful trick by using the defintion of the cosine and constructing an ordinary differential equation (ODE). First plug in to the definition of the characteristic function

$$\begin{aligned}
\Phi_X(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-\frac{x^2}{2}} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{itx} e^{-\frac{x^2}{2}} dx + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{itx} e^{-\frac{x^2}{2}} dx \\
&= -\frac{1}{\sqrt{2\pi}} \int_{\infty}^{0} e^{-itx} e^{-\frac{x^2}{2}} dx + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{itx} e^{-\frac{x^2}{2}} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} \left(e^{itx} + e^{-itx}\right) e^{-\frac{x^2}{2}} dx = \frac{\sqrt{2}}{\sqrt{\pi}} \int_{0}^{\infty} \cos(tx) e^{-\frac{x^2}{2}} dx
\end{aligned}$$

compute the dervative and integrate by parts

$$\begin{aligned}
\frac{d}{dt}\Phi_X(t) &= -\frac{\sqrt{2}}{\sqrt{\pi}} \int_{0}^{\infty} \sin(tx) x e^{-\frac{x^2}{2}} dx \\
&= \frac{\sqrt{2}}{\sqrt{\pi}} \sin(tx) e^{-\frac{x^2}{2}} \Big|_0^\infty - \frac{\sqrt{2}}{\sqrt{\pi}} \int_{0}^{\infty} t \cos(tx) e^{-\frac{x^2}{2}} dx \\
&= -t\Phi_X(t)
\end{aligned}$$

Thus we have the ODE $\frac{d}{dt}\Phi_X(t) = -t\Phi_X(t)$. The solution is given by

$$\Phi_X(t) = c\, e^{-\frac{t^2}{2}}, \tag{31}$$

with some integration constant c. In order to determine c we observe that $\Phi_X(0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = 1$ because we are dealing with a gaussian PDF, which has to integrate to one. So, using $\Phi_X(0) = 1$ in (31) gives $c = 1$. □

Generalisation to more than one dimension. The generalisation of definition 9 to l dimensions is straightforward by replacing tX by the scalarproduct. Let $X = (X_1, X_2, \ldots, X_l)$ and $\Phi(t) = \Phi(t_1, t_2, \ldots t_l)$

$$\Phi_X(t) := E_X\left(e^{i\langle X,t\rangle}\right) \tag{32}$$

8.1 Properties of the characteristic function

- **Hermitesch.**

$$\Phi_X(-t) = \overline{\Phi_X(t)} \tag{33}$$

- **Symmetry.** X is real-valued $\Leftrightarrow$ X is symmetric i.e., X and $-X$ have the same distribution.

- **Linear transformation**

$$\Phi_{aX+b}(t) = e^{itb}\Phi_{aX}(t) \tag{34}$$

- **Inversion.** If Φ_X is integrable, then the PDF may be reconstructed from the characteristic function

$$f_X(x) = \int_{-\infty}^{\infty} e^{-itx}\Phi_X(t)dt \tag{35}$$

- **Moment-generation.**

$$m_k := E(X^k) = \frac{1}{i^k}\frac{d^k}{dt^k}\Phi_X\Big|_{t=0} \tag{36}$$

- **Convolution.** For independent random variables X_1, X_2 the characteristic function of the sum $Y = X_1 + X_2$ is given by[2]

$$\Phi_Y(t) = \Phi_{X_1}(t)\Phi_{X_2}(t) \tag{37}$$

8.2 Relation to other generating functions

There are relationships between the characteristic function and other generating function such as the probability generating function, the moment generating function or the cumulant generating function.

[2] $E_Y(t) = E(e^{i(X_1+X_2)t}) = E(e^{iX_1t})E(e^{iX_2t}) = \Phi_{X_1}(t)\Phi_{X_1}(t)$

The probability generating function is defined over a $\mathbb{N}_0$ valued random variable X as $m_X(t) = E(t^X)$. Therefore we have the relationship $m_X(e^{it}) = \Phi_X(t)$.

The moment generating function of a real (or complex) random variable is defined as $M_X(t) = E(e^{tX})$. Therefore we have the relationship $M_{iX}(t) = M_X(it) = \Phi_X(t)$. This is of course only true if the moment generating function exists, which is not always the case in contrast to the characteristic function.

9 Probability distributions

9.1 The Γ function

The Γ function on the real axis is defined as

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \tag{38}$$

9.2 The χ^2 Distribution

We will follow http://www.real-statistics.com/chi-square-and-f-distributions/chi-square-distribution/chi-square-distribution-advanced/ here.

Motivation There are several arguments why the χ^2 distribution is important:

- The relation to hypothesis testing
- The χ^2 distribution is closely related to squared standard-normal random variables.

Let's have a look at the second point. The χ^2 distribution emerges from a sum of squared standard normal distributed random variables. Such squared random variables occur in estimations of sample-variances. As a consequence it will be an important distribution in estimating the goodness of a statistical model or confidence intervals based on empirical data. We will now derive as a motivation the χ^2 distribution in 1D.

Let Y be a random variable defined as $Y = X^2$, where $X \sim \mathcal{N}(0, 1)$. Since y is positive, we have for $y < 0$: $P(Y < y) = 0$. For $y \geq 0$ we have

$$\begin{aligned} P(Y < y) &= P(X^2 < y) = P(-\sqrt{y} < X < \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) = 2F_X(\sqrt{y}) - 1, \end{aligned}$$

where F_X is the CDF of $\mathcal{N}(0, 1)$ The PDF is obtained by taking the derivative

of $P(Y < y)$ with respect to y,

$$\begin{aligned}\frac{d}{dy}P(Y<y) &= \frac{2}{\sqrt{2\pi}}\frac{d}{dy}\int_{-\infty}^{\sqrt{y}} e^{-\frac{t^2}{2}}dt \\ &= \frac{\sqrt{2}}{\sqrt{\pi}}e^{-\frac{y}{2}}\frac{d}{dy}\sqrt{y} \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{y}{2}}y^{-\frac{1}{2}} \\ &= \frac{1}{\sqrt{2}\Gamma(\frac{1}{2})}e^{-\frac{y}{2}}y^{-\frac{1}{2}}\end{aligned}$$

In arbitrary dimensions the χ^2 distribution assumes is defined as:

Definition 10. *The PDF of the χ^2 distribution is given by:*

$$f(x;k) = \begin{cases} \frac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}, & x > 0 \\ 0 & else \end{cases} \tag{39}$$

Where the parameter k is called degrees of freedom.

Theorem 8. *The moment generating function of the χ^2 distribution is given by*

$$M_x(t) = (1-2t)^{-\frac{k}{2}}$$

Proof.

$$M_x(t) = \int_0^\infty e^{tx}f(x)dx = \int_0^\infty \frac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}(1-2t)}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}dx.$$

Substituting $y = \frac{x}{2}(1-2t)$, we get

$$M_x(t) = \frac{(1-2t)^{-\frac{k}{2}}}{\Gamma(\frac{k}{2})}\int_0^\infty y^{\frac{k}{2}-1}e^{-y}dy = (1-2t)^{-\frac{k}{2}},$$

where we have used that $\Gamma(k/2)$ equals the integral (by definition). □

With this result we can easily calculate mean and variance via the first moments $m_i,\ i = 1, 2$ for the χ^2 distribution,

$$m_1 = \left.\frac{dM_x(t)}{dt}\right|_{t=0} = k \quad \Rightarrow \mu = m_1 = k \tag{40}$$

$$m_2 = \left.\frac{d^2M_x(t)}{dt^2}\right|_{t=0} = k(k+2) \quad \Rightarrow \sigma^2 = m_2 - \mu^2 = 2k \tag{41}$$

We have used for the variance $\sigma^2 = E((X - E(X))^2) = E(X^2) - E(X)^2$. Alternatively - and prettier - one could have derived directly the generating function for the cumulants, $g_X(t) = \ln M_X(t)$. Furhtermore we have an additivity property. Let x have $\chi^2(m)$ and y have $\chi^2(n)$ distribution. Then $x+y$ has a $\chi^2(m+n)$ distribution. We calculate the MGF

$$M_{x+y}(t) = M_x(t)M_y(t) = (1-2t)^{-\frac{m}{2}}(1-2t)^{-\frac{n}{2}} = (1-2t)^{-\frac{m+n}{2}}. \tag{42}$$

But this is the MGF of $\chi^2(m+n)$. Using the argument that the MGF uniquely determines the PDF this shows the claim.

Theorem 9. *Let $X_n \sim \mathcal{N}(0,1)$ be a standard normal distributions of k stochastically independent random variables. Then the sum of their squares distribution*

$$w = \sum_{n=1}^{k} X_n^2,$$

obey a χ_k^2 distribution.

Proof. We will argue via the moment generating function and use the fact that the MGF uniquely determines the PDF.

$$M_w(t) = M_{x_1^2+\cdots+x_k^2}(t) = M_{x_1^2}(t) \dots M_{x_k^2}(t) = (M_{x^2}(t))^k, \tag{43}$$

which factorizes because the variables are independent.Thus, we need to calculate the MGF of x^2 with $X \sim \mathcal{N}(0,1)$.

$$\begin{aligned} M_{x^2}(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx^2} e^{-\frac{x^2}{2}} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}(1-2t)} dx \\ &= (1-2t)^{-\frac{1}{2}} \end{aligned}$$

where we have substituted $y = x\sqrt{1-2t}$ and employed the property of $\mathcal{N}(0,1)$ that $1/\sqrt{2\pi} \int e^{-y^2/2} dy = 1$. Thus we have for the MGF

$$M_w(t) = (M_{x^2}(t))^k = (1-2t)^{-\frac{k}{2}} \tag{44}$$

On the other hand, according to Theorem 8 this is the MGF of the χ^2 distribution. Since the MGF uniquely determines the PDF it follows that w has distribution χ^2. □

- Moment generating functions for $f(x)$, eg: x^2
- χ^2 plots
- χ^2 distance
- χ^2 test

References

[1] J. H. Curtiss. A note on the theory of moment generating functions. *Ann. Math. Statist.*, 13(4):430–433, 12 1942.

[2] Trevor J. Hastie, Robert John Tibshirani, and Jerome H. Friedman. *The elements of statistical learning : data mining, inference, and prediction.* Springer series in statistics. Springer, New York, 2009. Autres impressions : 2011 (corr.), 2013 (7e corr.).

[3] Zhexue Huang. Clustering large data sets with mixed numeric and categorical values. In *In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 21–34, 1997.

[4] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.*, 2(3):283–304, September 1998.

[5] WH Kruskal and WA Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, pages 583–621, 1952.

[6] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 911–916, Washington, DC, USA, 2010. IEEE Computer Society.

[7] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a dataset via the gap statistic. 63:411–423, 2000.

[8] Ulrike von Luxburg. *Statistical Learning with Similarity and Dissimilarity Functions*. PhD thesis, 2004.