Engineering and Applied Science Programs for Professionals
Whiting School of Engineering
Johns Hopkins University
685.621 Algorithms for Data Science
Probabilities

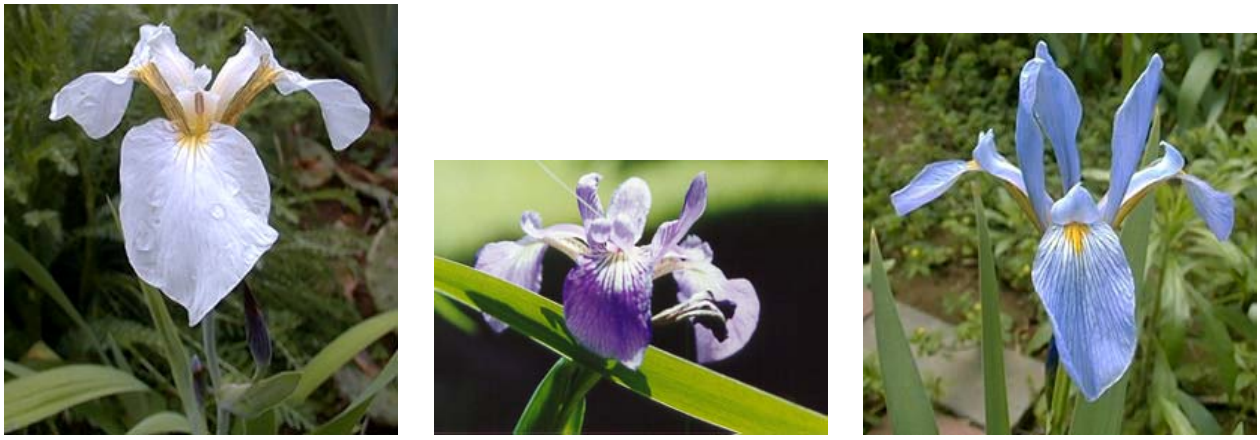
This document provides a rollup of the probability methods covered in this module. The following list of topics is covered in this module allowing for an understanding of how we will analyze, design and develop probability methods for Machine Learning.

Contents

1 Probabilities

When analyzing data an intuitive reaction is to graph the data in its raw format. This will allow you to see two or three dimensions at a time. This of course assumes our data consists of numerical values or a representation that can be visually represented. This has its limitations and in many cases is not a practical approach. When data is not represented in numerical form there is some preprocessing involved either prior to using an algorithm or within the algorithm itself. In this document the assumption is that data is purely represented by some real numerical values, it is complete and within allowable ranges, e.g., $\mathbf{x} = [x_1, x_2, \cdots, x_n]$ where $\mathbf{x} \in \mathbb{R}$. When data is analyzed using probability theory it is possible to use pattern recognition techniques as well as machine learning to analyze the data in order to determine what the data represents.

Data of other forms and how to represent them will be discussed in the Data Processing module. Let's consider the Iris data set containing 3 classes of 50 instances each, where each class refers to a type of Iris plant. Five attributes/features were collected for each plant instance.



(a) Setosa (b) Versicolor (c) Virginica

Figure 1: Iris plant images courtesy of STPRTool

Let's look at the three different flowers and generate some probability values shown in Table 1.

From this complex table there is sufficient information to make determinations as to the data and how to best process the data. Now let's see what each of the test statistics means without evaluating the Plant Class for now. Let's consider the minimum and maximum values. In this case it can be determined that the data has the range of approximately 0.1 to 7.9 or mathematically represented as $\mathbf{x} \in [0.1, 7.9]$ in which case it will be easier to represent the range as $\mathbf{x} \in [0, 8]$. If we break this down even farther each of the features can have their own ranges as shown in in Table 2. In this case the ranges are in units of centimeters and are not that wide spread. In the data processing module data normalization will be discussed to ensure various statistical measures give the correct representation.

Now let's look at the mean and standard deviation of the data to get a representation of the data and what is meant to ask what is the "mean" and "standard deviation" of the data. Let's consider the data by features for all flower types, in this case the mean and corresponding standard deviation are shown in Table 3. Analyzing the results several conclusion can be drawn. First the largest mean is in the sepal length where the mean is 5.8433 with a corresponding standard deviation of 0.8281 and the smallest mean is the petal width with a mean of 1.1987 and corresponding standard deviation of 0.7632. Looking at the results of two features it may be concluded that the sepal length has a tighter distribution in the data. If the data is investigated further and the minimum and maximum feature values are included it is noted that the spread of data is $7.9 - 4.3 = 3.6$ for the sepal length and the spread for the petal width is $2.5 - 0.1 = 2.4$ which leads to an initial analysis that the larger mean with smaller

Table 1: Data Analysis Statistics

Test Statistics	Flower Type	Sepal Length	Sepal Width	Petal Length	Petal Width	Plant Class
Minimum	Setosa	4.3000	2.3000	1.0000	0.1000	1
	Versicolor	4.9000	2.0000	3.0000	1.0000	2
	Virginica	4.9000	2.2000	4.5000	1.4000	3
Maximum	Setosa	5.8000	4.4000	1.9000	0.6000	1
	Versicolor	7.0000	3.4000	5.1000	1.8000	2
	Virginica	7.9000	3.8000	6.9000	2.5000	3
Mean	Setosa	5.0060	3.4180	1.4640	0.2440	1
	Versicolor	5.9360	2.7700	4.2600	1.3260	2
	Virginica	6.5880	2.9740	5.5520	2.0260	3
Standard Deviation +/-	Setosa	0.3525	0.3810	0.1735	0.1072	1
	Versicolor	0.5162	0.3138	0.4699	0.1978	2
	Virginica	0.6359	0.3225	0.5519	0.2747	3
Skewness	Setosa	0.1165	0.1038	0.0697	1.1610	1
	Versicolor	0.1022	-0.3519	-0.5882	-0.0302	2
	Virginica	0.1144	0.3549	0.5328	-0.1256	3
Kurtosis	Setosa	2.6542	3.6851	3.8137	4.2965	1
	Versicolor	2.4012	2.5517	2.9256	2.5122	2
	Virginica	2.9121	3.5198	2.7435	2.3387	3

Table 2: Data Minimum and Maximum Values by Feature

Test Statistics	Sepal Length	Sepal Width	Petal Length	Petal Width
Minimun	4.3000	2.000	1.000	0.1000
Maximum	7.9000	4.4000	6.9000	2.5000

standard deviation is not necessarily the best feature for evaluation.

This leads to the question as too which is the best feature(s) in the data for evaluation? It is not trivial to answer this question, meaning, what is it that is being evaluated in the data. Let's assume that currently we are interested in evaluating the four features and how their "test statistics" measure. So far only the minimum, maximum, mean and standard deviation of the data have been evaluated without regard to the plant class which the full set of test statistics are shown in Table 1. To complete the initial analysis let's consider the skewness and kurtosis by feature and not flower type. The results in Table 4.

Let's look at the meaning of skewness and determine the feature(s) with the "best" skewness value. The measure of data asymmetry around the sample mean is know as skewness. In the event the skewness value is negative, the data has a spread to the left of the mean. When the value is positive the spread is to the left of the mean. The value returned is zero when perfect symmetry is about the mean. With this said the feature with the best skewness is the petal width of the Iris flower with a value of -0.1039.

Table 3: Mean and Standard Deviation by Feature

Test Statistics	**Sepal Length**	**Sepal Width**	**Petal Length**	**Petal Width**
Mean	5.8433	3.0540	3.7587	1.1987
Standard Deviation	0.8281	0.4336	1.7644	0.7632

Table 4: Skewness and Kurtosis by Feature

Test Statistics	**Sepal Length**	**Sepal Width**	**Petal Length**	**Petal Width**
Skewness	0.3118	0.3307	-0.2717	-0.1039
Kurtosis	2.4264	3.2414	1.6046	1.6648

Now let's look at the meaning of the kurtosis and determine the feature(s) with the "best" kurtosis value. So in this case the kurtosis provides a measure of how prone is the data to outliers. The introduction to outliers in data is provided in the data preprocessing module so for now outlier(s) is simply defined as a data point that differs significantly from other data points for a particular feature. A kurtosis value of 3 indicates the data has a normal distribution. Values that are above 3 are known to have outliers and values that are less than 3 do not have outliers. This can be normalized to have a mean value of 0 so anything above 0 will have outliers and below 0 do not have outliers. So in analyzing the results from Table 4, it can be seen that the sepal width has the value closest to 3.

This analysis provides an initial insight of the data. From the analysis of the test statistics used it is not clear which feature(s) are the best or an ability to rank them in some form at this time. These statistics do give us insight into what the features represent in terms of the range of the data, the mean and the spread of the data. In the upcoming data processing module additional preprocessing methods will be introduced for the purpose of pattern recognition and pattern classification. In addition it should be mentioned that the statistics in this section are best suited for data that is normally distributed.

2 Distributions

This section is from (Bishop, 1995).

Let's assume the knowledge of random variables is known for the sake of analyzing data and describing data distributions.

One of the most straight forward approaches to density estimation is to represent the probability density $p(\mathbf{x})$ in terms of a specific functional form which contains a number of adjustable parameters. The values of the parameters can then be optimized to give the best fit tot he data. The simplest and most widely used parametric model is the normal or Gaussian distribution, which has a number of convenient analytical and statistical properties. Since our aim is to explain the basic principles of parametric density estimation, we shall limit our discussion to normal distribution.

We shall also describe the two principal techniques for determining the parameters of the model distribution, known respectively as maximum likelihood and Bayesian inference.

Normal Distribution, the normal density function, for the case of a single variable, can be written in the following form

$$p(x) = \frac{1}{2\pi\sigma^2} \exp\left\{ - \frac{(x-\mu)^2}{2\sigma^2} \right\} \tag{1}$$

where μ and σ^2 are the mean and the variance respectively, and the parameter σ (which is the square root of the

variance) is called the *standard deviation*. The coefficient in front of the exponential in Equation (1) ensures that $\int_{-\infty}^{\infty} p(x)dx = 1$. The mean and variance of the one-dimensional normal distribution satisfy the following

$$\mu(\mathbf{x}) = E[\mathbf{x}] = \int_{-\infty}^{\infty} xp(x)dx \tag{2}$$

$$\sigma^2(\mathbf{x}) = E[(\mathbf{x} - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx \tag{3}$$

where $E[\cdot]$ denotes the expectation.

In d dimensions the general multivariate normal probability density can be written as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \Big\} \tag{4}$$

where the mean $\boldsymbol{\mu}$ is now a d-dimensional vector, $\mathbf{\Sigma}$ is a $d \times d$ *covariance matrix*, and $|\mathbf{\Sigma}|$ is the determinate of $\mathbf{\Sigma}$. The pre-factor in Equation (4) ensures that $\int_{-\infty}^{\infty} p(\mathbf{x})d\mathbf{x} = 1$. The density function $p(\mathbf{x})$ is governed by the parameters $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$, which satisfies

$$\boldsymbol{\mu} = E[\mathbf{x}] \tag{5}$$

$$\mathbf{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]. \tag{6}$$

From Equation (6) we see that $\mathbf{\Sigma}$ is a symmetric matrix, and therefore has $d(d+1)/2$ independent components. There are also d independent elements in $\boldsymbol{\mu}$, and so the density function is completely specified once the values of $d(d+3)/2$ parameters have been determined. The quantity

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \tag{7}$$

which appears in the exponent in Equation (4), is called the *Mahalanobis* distance from $\mathbf{x}$ to $\boldsymbol{\mu}$. The surface of constant probability density for Equation (4) are hyperellipsoids on which Δ^2 is constant, as shown for the case of two dimensions in Figure 2. The principal axes of the hyperellipsoids are given by the eigenvectors $\mathbf{u}_i$ of $\mathbf{\Sigma}$ which satisfy

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{8}$$

and the corresponding eigenvalues λ_i give the variances along the respective principal directions.

It is sometimes convenient to consider a simplified form of Gaussian distribution in which the covariance matrix is diagonal,

$$(\mathbf{\Sigma})_{ij} = \delta_{ij}\sigma_j^2 \tag{9}$$

which reduces the total number of independent parameters in the distribution to $2d$. In this case the contours of constant density are hyperellipsoids with the principal directions aligned with the coordinate axes. The components of $\mathbf{x}$ can be written as the product of the distributions for each of the components separately in the form

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i). \tag{10}$$

Further simplification can be obtained by choosing $\sigma_j = \sigma$ for all j, which reduces the number of parameters still further to $d + 1$. The contours of constant density are then hypersheres.

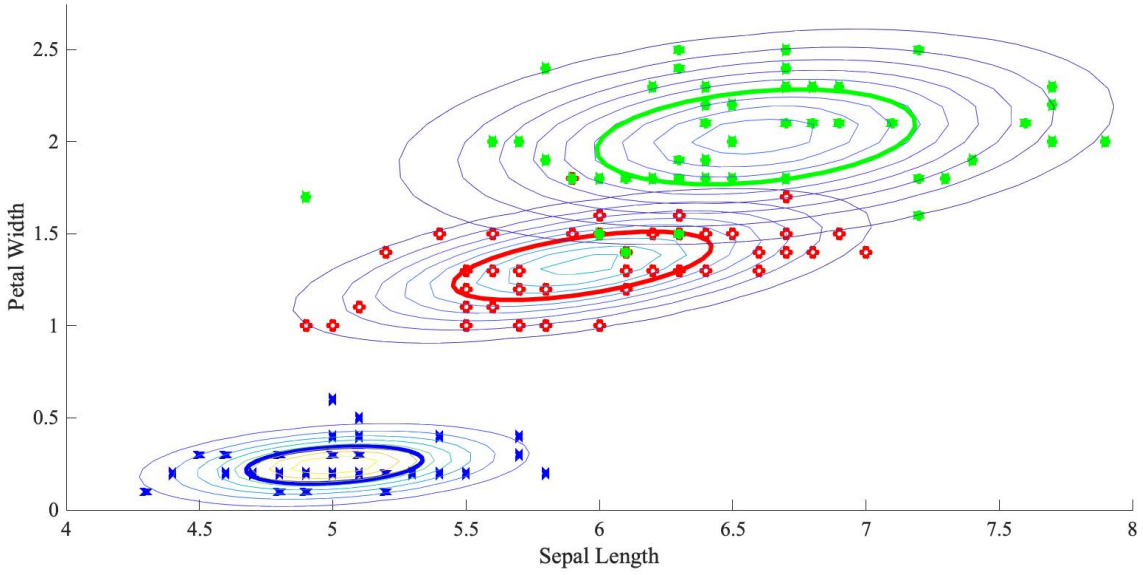


Figure 2: Normal distributions for the sepal length vs the petal width classes.

2.1 Moments of Distributions

When a set of data values has a strong central tendency, that is, a tendency to cluster around some particular value, then it may be useful to characterize the data set by a few values that have been calculated numerically, which are known as the data moments. This also assumes that the data has a random nature and is limited by a set range.

In this section, the moments described are the mean, standard deviation, skewness, and kurtosis. In addition the data minimum and maximum are introduced as they are important in the analysis of data. Table 5 shows the mathematical representation of the moments and the inputs variable $\mathbf{x} = [x_1, x_2, \cdots, x_n]$ where $\mathbf{x} \in \Re$.

Table 5: Data Analysis Statistics

Test Statistics	**Statistical Function $F(\cdot)$**
Minimum	$F_{\min}(\mathbf{x}) = \min(\mathbf{x}) = x_{min}$
Maximum	$F_{\max}(\mathbf{x}) = \max(\mathbf{x}) = x_{max}$
Mean	$F_{\mu}(\mathbf{x}) = \mu(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$
Standard Deviation	$F_{\sigma}(\mathbf{x}) = \sigma(\mathbf{x}) = \left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu(\mathbf{x})\right)^2\right)^{1/2}$
Skewness	$F_{\gamma}(\mathbf{x}) = \gamma(\mathbf{x}) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu(\mathbf{x})\right)^3}{\sigma(\mathbf{x})^3}$
Kurtosis	$F_{\kappa}(\mathbf{x}) = \kappa(\mathbf{x}) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu(\mathbf{x})\right)^4}{\sigma(\mathbf{x})^4}$

References

[1] Bishop, Christopher M., *Neural Networks for pattern Recognition*, Oxford University Press, 1995

[2] Bishop, Christopher M., *Pattern Recognition and Machine Learning*, Springer, 2006

[3] Cormen, Thomas H., Leiserson, Charles E., Rivest, Ronal L., and Stein, Clifford, *Introduction to Algorithms*, 3rd Edition, MIT Press, 2009

[4] Duin, Robert P.W., Tax, David and Pekalska, Elzbieta, *PRTools*, http://prtools.tudelft.nl/

[5] Franc, Vojtech and Hlavac, Vaclav, *Statistical Pattern Recognition Toolbox*, https://cmp.felk.cvut.cz/cmp/software/stprtool/index.html

[6] Fukunaga, Keinosuke, *Introduction to Statistical Pattern Recognition*, Academic Press, 1972

[7] Machine Learning at Waikato University, *WEKA*, https://www.cs.waikato.ac.nz/ ml/index.html

[8] Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P., *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Jan 31, 1986

[9] Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P., *Numerical Recipes: The Art of Scientific Computing*, 3rd Edition, Cambridge University Press, September 10, 2007

[10] Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P., *Numerical Recipes: The Art of Scientific Computing*, 3rd Edition, http://numerical.recipes/

[11] Press, William H., *Opinionated Lessons in Statistics*, http://www.opinionatedlessons.org/