Symbols, Patterns and, Signals Notes

by Tyler Wright

github.com/Fluxanoia fluxanoia.co.uk

These notes are not necessarily correct, consistent, representative of the course as it stands today, or rigorous. Any result of the above is not the author's fault.

These notes are marked as unsupported, they were supported up until June 2020.

These notes are incomplete and will remain so for the foreseeable future.

Contents

1 Data Acquisition

1.1 Analogue to Digital Conversion

There are two steps to this conversion, sampling and quantisation. They can be done in any order.

1.1.1 Nyquist-Shannon Sampling Theorem

If a function f contains no frequencies higher than some $h_{\max}$ hertz, it is completely determined by sampling at points spaced $\frac{1}{2 \cdot h_{\max}}$ apart.

2 Data Characteristics

2.1 Measures of Distance

A valid distance measure $D : A \times A \to \mathbb{R}$ for some data set A has the following properties, it is:

- non-negative,
- reflexive ($D(a, b) = 0 \iff a = b$),
- symmetric,
- satisfies the triangle inequality ($D(a, b) + D(b, c) \geq D(a, c)$).

2.1.1 Euclidean Distance in $\mathbb{R}^n$ (p-norm distance)

For two vectors x and y in $\mathbb{R}^n$, we have the Euclidean distance D is:

$$D(x, y) := \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}.$$

2.1.2 Chebyshev Distance in $\mathbb{R}^n$ (∞-norm distance)

For two vectors x and y in $\mathbb{R}^n$, we have the Chebyshev distance D is:

$$D(x, y) := \lim_{n \to \infty} \left(\sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \right) = \max_{i \in [n]} (|x_i - y_i|).$$

2.1.3 Time Series Distance

Finding the distance between two time series x and y of length n and m (resp.) can be found using Dynamic Time Warping:

$$D_{tw}(x, y) := D(x_1, y_1) + \min\{D_{tw}(x, y'), D_{tw}(x', y), D_{tw}(x', y')\},$$

where D is some numerical distance measure and x' and y' are the time series length $n-1$ and $m-1$ (resp.) corresponding to x and y with the first element removed.

2.1.4 Hamming Distance

When given two strings of the same length, the Hamming distance between them is how many characters differ in the strings at each index.

2.1.5 Edit Distance

When given two strings of any length, the edit distance between them is the smallest number of insertions, substitutions and, deletions that can transform one string into the other (or vice versa).

2.1.6 Wu and Palmer Distance

This measure is based on a hierarchy of word semantics, a graph of relationships between words based on meaning. Using the shortest distance between the words d_1 and the shortest distance from a most specific ancestor to the path d_2 we have the distance measure:

$$D(w_1, w_2) := \frac{2 \cdot d_2}{d_1 + 2 \cdot d_2} - 1.$$

2.2 Summary Statistics

2.2.1 Mean

We take $X = \{x_1, \ldots, x_n\}$ to be a data set. The mean $\bar{X}$ is defined as follows:

$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} x_i.$$

2.2.2 Standard Deviation and Variance

We take $X = \{x_1, \ldots, x_n\}$ to be a data set. We have that for σ_X, the standard deviation of X, the variance of X is σ_X^2. We define the variance (and thus the standard deviation) as follows:

$$\sigma_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2.$$

2.2.3 Covariance

We take $X = \{x_1, \ldots, x_n\}$ to be a data set consisting of m-dimensional row vectors. We define the covariance matrix:

$$\Sigma = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^T (x_i - \mu),$$

where μ is the m-dimensional row vector where the j^{th} entry corresponds to the mean of the j^{th} entry of each x_i. This yields a $m \times m$ matrix that is square and symmetric with the variance of the j^{th} entry of each x_i on the j^{th} value on the diagonal.

Eigenvalues and Eigenvectors As the matrix is symmetric we can diagonalise it and find the eigenvalues and eigenvectors. The major axis is the eigenvector corresponding to the largest eigenvalue and the minor axis is the eigenvector corresponding to the smallest value.

2.2.4 Data Normalisation

Data may need to be normalised before we use our distance measures on it. We consider a data set $X = \{x_1, \ldots, x_n\}$ with mean μ and standard deviation σ.

Scaling We map each x_i for $i \in [n]$ as follows:

$$x_i \mapsto \frac{x_i - \min(X)}{\max(X) - \min(X)}.$$

Standardisation We map each x_i for $i \in [n]$ as follows:

$$x_i \mapsto \frac{x_i - \mu}{\sigma}.$$

Scaling to Unit Length We map each x_i for $i \in [n]$ as follows:

$$x_i \mapsto \frac{x_i}{|x_i|}.$$

where $|x_i|$ denotes the magnitude of x_i.

2.2.5 Outliers

A small amount of values significantly different to the remainder of the data set.

3 Data Modelling

A model is a description of data, it should generalise, either as an abstraction or a simplification.We can quantify the performance of a model by how well it maps the data to the desired solution.

3.1 Model Parameters

Models are defined in terms of parameters which could be obtained through trial and error or training data (through tuning or training the model).

3.2 Overfitting

Training a model too hard on a specific data set can cause it to 'overfit'. This means it performs very well on the trained data set but does not generalise well.

This can happen in a variety of cases, some including:

- There is too little data,
- There is too little data representing some key part of the data distribution,
- The function class is too complex.

3.2.1 Cross-validation

We can decide to train our data on only a fraction of our data set and then use the remainder to test whether our data is overfitted. This can be repeated multiple times with different subsets of the original data set. However, this can fail on smaller data sets or ones with many parameters.

We can combine cross-validation with regularisation to produce regularising matrices to penalise overfitted parameters.

3.3 Bayesian Fitting

Suppose we are interested in some weighting vector $\mathbf{w}$ of a distribution. We consider the prior $\mathbb{P}(\mathbf{w})$ and posterior $\mathbb{P}(\mathbf{w} \mid \mathbf{X})$ probability of our vector given some sample data $\mathbf{X}, \mathbf{y}$ and relate them using Bayes' theorem:

$$\mathbb{P}(\mathbf{w} \mid \mathbf{X}) = \frac{\mathbb{P}(\mathbf{X} \mid \mathbf{w})\mathbb{P}(\mathbf{w})}{\mathbb{P}(\mathbf{X})}.$$

By taking the natural logarithm, we see that:

$$\ln(\mathbb{P}(\mathbf{w} \mid \mathbf{X})) = \ln(\mathbb{P}(\mathbf{X} \mid \mathbf{w})) + \ln(\mathbb{P}(\mathbf{w})) + \ln\left(\frac{1}{\mathbb{P}(\mathbf{X})}\right).$$

We consider a fixed data set, so $\mathbf{X}$ is not a variable. Assuming we are sampling from the multivariate normal, we can calculate $\ln(\mathbb{P}(\mathbf{w} \mid \mathbf{X}))$ in terms of the equation above and in terms of the multivariate normal distribution. Equating these allows us to identify equations for the regularised, linear regression mean and variance parameters:

$$\begin{aligned}\hat{\mu} &= ((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} + \sigma^2 \mathrm{Var}(\mathbf{w})^{-1})^{-1}\mathbf{X}\mathbf{y}, \\ \hat{\Sigma} &= \sigma^2((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} + \sigma^2 \mathrm{Var}(\mathbf{w})^{-1})^{-1}.\end{aligned}$$

3.4 Deterministic Models

Deterministic models produce an output without a measure of confidence in that output.

3.4.1 Single Variable Linear Regression

For a two dimensional set of data:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

We calculate a gradient r and an intercept c to uniquely define the regression line ($y = rx + c$) on D:

$$\begin{aligned}r &= \frac{\sum_{i=1}^{n}(x_i \cdot y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^{n}(x_i^2) - n\bar{x}^2}, \\ c &= \bar{y} - r \cdot \bar{x}.\end{aligned}$$

Outliers have disproportionate effects due to the squares used by the measure.

For a two dimensional set of data:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

where each x_i is k-dimensional, we can use matrices to calculate our coefficients:

$$R = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y,$$

where:

$$R = \begin{pmatrix} r_0 \\ r_1 \\ \vdots \\ r_k \end{pmatrix} \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & \leftarrow & x_1 & \rightarrow \\ 1 & \leftarrow & x_2 & \rightarrow \\ \vdots & & \vdots & \\ 1 & \leftarrow & x_n & \rightarrow \end{pmatrix}.$$

Instead of using k-dimensional data, we can instead use powers of each x^i to form fitted polynomials. In this case, we have a least squares linear regression line $y = r_0 + r_1 x^1 + \cdots + r_n x^n$.

3.4.2 Multivariate Linear Regression

We consider a set of data with k data points, with the i^{th} data point corresponding to $(\mathbf{x}_i, \mathbf{y}_i)$ the n and m-dimensional, input and output vectors (resp.). We assume each $\mathbf{y}_i$ is multivariate normal, conditional on its corresponding $\mathbf{x}_i$ with some n-dimensional weight vector $\mathbf{w}$:

$$\mathbb{P}(\mathbf{y}_i \,|\, \mathbf{x}_i, \mathbf{w}) = f(\mathbf{y}_i),$$

where f is the multivariate normal probability density function with mean $\mathbf{x}_i \cdot \mathbf{w}$ and variance σ^2. We can expand this to the full data set with:

$$\mathbf{y} = \begin{pmatrix} \leftarrow & \mathbf{y}_1 & \rightarrow \\ \leftarrow & \mathbf{y}_2 & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{y}_n & \rightarrow \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} \leftarrow & \mathbf{x}_1 & \rightarrow \\ \leftarrow & \mathbf{x}_2 & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n & \rightarrow \end{pmatrix}.$$

giving us:

$$\mathbb{P}(\mathbf{y} \,|\, \mathbf{X}, \mathbf{w}) = f(\mathbf{y}), \tag{1}$$

where f is the multivariate normal probability density function with mean $\mathbf{X} \cdot \mathbf{w}$ and variance $\sigma^2 I$. By using maximum likelihood estimation, we can see that our optimal weight vector $\hat{\mathbf{w}}$ is equal to $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

3.4.3 Regularisation and Multivariate Linear Regression

In an effort to penalise exceptionally large and potentially overfitted regression weights, we use regularisation. We modify the prediction of the distribution of $\mathbf{y}$ seen in (1):

$$\mathbb{P}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = f(\mathbf{y}) \left[e^{-\frac{1}{2}\mathbf{w}^{\mathrm{T}}\Lambda\mathbf{w}} \right],$$

with diagonal regularisation matrix Λ so when we take the natural logarithm:

$$\ln(\mathbb{P}(\mathbf{y} \mid \mathbf{X}, \mathbf{w})) = \ln(f(\mathbf{y})) - \frac{1}{2}\mathbf{w}^{\mathrm{T}}\Lambda\mathbf{w}.$$

Note that where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $\mathbf{w} = (w_1, \ldots, w_n)$:

$$\mathbf{w}^{\mathrm{T}}\Lambda\mathbf{w} = \sum_{i=1}^{n} \lambda_i \cdot w_i^2.$$

We have the maximum likelihood estimate in this case is:

$$\hat{\mathbf{w}} = ((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} + \sigma^2\Lambda)^{-1}\mathbf{X}\mathbf{y}.$$

3.5 Probabilistic Models

Probabilistic models pair outputs with measures of confidence, probability.

3.5.1 Maximum Likelihood Estimation

We first define the argmax function:

$$\mathrm{argmax}_\theta(f(\theta)),$$

for some objective function f, is the θ such that $f(\theta)$ is maximised - denoted by $\hat{\theta}$. We define argmin similarly, and can see that:

$$\begin{aligned} \mathrm{argmax}_\theta(f(\theta)) &= \mathrm{argmax}_\theta(\ln(f(\theta))) \\ &= \mathrm{argmin}_\theta(-\ln(f(\theta))), \end{aligned}$$

as ln is monotone increasing.

When considering some data set D with parameters encapsulated by θ, we take $\mathbb{P}(D \mid \theta)$ to be the function returning the probability of D occuring based on the parameters of θ. Thus:

$$\mathrm{argmax}_\theta(\mathbb{P}(D \mid \theta)),$$

represents the parameter values that are most likely to produce this data set. Practically, this can be carried out by taking the derivative of $\ln(\mathbb{P}(D \mid \theta))$ and taking derivatives of it with respect to the parameters. We set the derivatives to zero and solve for θ.

3.5.2 Approximate Bayesian Computation

If we have a very small amount of data, we can take simulations of the (suspected) underlying distributions varying across its parameters and identify the simulations which contain our small data set. We can plot the histogram of these supersets to see which parameters are most likely to produce our small data set.

However, in practice - this doesn't work very well.

3.5.3 Maximum a Posteriori

If we have some prior information about the distribution of θ, we can consider assimilating that into our estimation. We take the posterior to be:

$$\mathbb{P}(\theta \mid D) = \frac{\mathbb{P}(D \mid \theta)\mathbb{P}(\theta)}{Z},$$

for some normalising term Z. Note that, $\mathbb{P}(D \mid \theta)$ is our likelihood and $\mathbb{P}(\theta)$ is our prior. Thus, we extend our Maximum Likelihood Estimator to use the posterior instead, giving us the Maximum a Posteriori estimator:

$$\hat{\theta} = \text{argmax}_\theta(\mathbb{P}(\theta \mid D)).$$

4 Data Representation

Data representation is about extracting features from data.

4.1 Preprocessing

We perfom preprocessing to remove noise, rectify missing values and, remove redundancies and inconsistencies. We do this by:

- Data cleaning, removing noise and inconsistent data,
- Data integration, combining data from multiple sources,
- Data selection, filtering and extracting the desired data,

allowing us to transform the data so it is ready for analysis.

4.2 Digital Signal Processing

Digital signal processing is about processing and manipulating signals using digital techniques.

4.2.1 Shannon's Sampling Theorem

An analogue signal of frequency at most $f_{\max}$ should be sampled at $2 \cdot f_{\max}$ to stop aliasing of the reconstructed signal.

4.2.2 Quantisation

Quantisation levels are the values you are assigning your input signal to in order to represent it digitally.

4.3 Fourier Series

For some periodic function $f : \mathbb{R} \to \mathbb{R}$ with period T, we can write f as an infinite sum:

$$f(x) = \sum_{n=0}^{\infty} a_n \cos\left(\frac{2\pi n}{T}x\right) + b_n \sin\left(\frac{2\pi n}{T}x\right),$$

where $(a_n)_{n\in\mathbb{Z}_{\geq 0}}$ and $(b_n)_{n\in\mathbb{Z}_{\geq 0}}$ are sequences in $\mathbb{R}$. Also, with some consideration of the properties of sin and cos, we can write this as:

$$f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{2\pi n}{T}x\right) + b_n \sin\left(\frac{2\pi n}{T}x\right).$$

The solution for these coefficients is as follows:

$$a_k = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(x)\cdot\cos\left(\frac{2\pi n}{T}x\right)dx$$
$$b_k = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(x)\cdot\sin\left(\frac{2\pi n}{T}x\right)dx,$$

or taking λ to be our frequency ($\lambda = \frac{1}{T}$):

$$a_k = 2\lambda\int_{-\frac{T}{2}}^{\frac{T}{2}} f(x)\cdot\cos\left(2\pi n\lambda x\right)dx$$
$$b_k = 2\lambda\int_{-\frac{T}{2}}^{\frac{T}{2}} f(x)\cdot\sin\left(2\pi n\lambda x\right)dx.$$

4.3.1 Fourier Transform

We can apply the Fourier series process to non-periodic functions too, in that case yielding a Fourier transform. In the case of some $f : \mathbb{R} \to \mathbb{R}$ (not necessarily periodic):

$$F(\lambda) = \int_{-\infty}^{\infty} f(x)e^{-2\pi\lambda ix}\,dx,$$

is the Fourier transform of f, where $i = \sqrt{-1}$. We can also inverse this process:

$$f(x) = \int_{-\infty}^{\infty} F(\lambda)e^{2\pi\lambda ix}\,du.$$

For $f : [n] \to [n]$ (discrete), we have the corresponding pair:

$$F(\lambda) = \frac{1}{n+1}\sum_{x=0}^{n} f(x)e^{-\frac{2\pi\lambda i}{n+1}x}, \qquad f(x) = \sum_{\lambda=0}^{n} F(\lambda)e^{\frac{2\pi\lambda i}{n+1}x}.$$

By using Euler's identity, we can expand our expressions using e into the sum of sin and cos. Furthermore, we can consider taking the real R and imaginary I parts of F to derive quantities:

$$\begin{aligned} F(\lambda) &= R(\lambda) + i\cdot I(\lambda) \\ \text{Magnitude, } |F(\lambda)| &= \sqrt{R^2(\lambda) + I^2(\lambda)} \\ \text{Phase angle, } \phi(\lambda) &= \tan^{-1}\left(\tfrac{I(\lambda)}{R(\lambda)}\right) \\ \text{Polar coordinate of } F(\lambda) &= |F(\lambda)|e^{i\phi(\lambda)} \end{aligned}$$

5 Data Visualisation

5.1 Scatter Plots

Scatter plots are useful for visualising two dimensional data and their relation. When we get up to higher dimensions, we can represent the data as a matrix of scatter plots where we choose two of the possible parameters at a time.

5.2 Histograms

For discrete data, this is a bar chart. For continuous data, we sort data into bins of some non-zero width.

5.3 Box Plots

Box plots give a good idea of some of the key values of distributions like the interquartile range, median and, range.

5.4 Surfaces

We can use 3D surfaces to represent 3D data but also, we can encode a fourth dimension into the colour of the plot at each point.

5.5 The Lie Factor

We can quantify how much we are over/under-stating effects in our data with the Lie Factor (LF):

$$\text{LF} = \frac{\text{size of effect in visualisation}}{\text{size of effect in data}}.$$

So, an LF greater than one means we are over-stating and an LF less than one means we are under-stating.