ML Preparation Notes

March 24, 2021

1 Basics

The *correlation* between two sets of data is a measure of the strength of the relationship between them. In particular, Pearson's correlation coefficient is a measure of linear relationship between two sets of data. Let X and Y be two random variables. Then Pearson's correlation coefficient $\rho(X, Y)$ is defined as:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \, \sigma_Y} \tag{1}$$

Two important facts about the Pearson's correlation coefficient [2]:

1. $-1 \leq \rho(X, Y) \leq 1$
2. $|\rho(X, Y)| = 1$ iff there exists $a \neq 0$ and b such that $Y = aX + b$.

2 Timeseries

The material in the section follows [1]. A *time series* is a sequence of random variables $\{X_1, X_2, \ldots\}$. A complete model of such a time series would require specifying the joint probability distributions of all random vectors $(X_1, \ldots, X_n)'$ for all $n \geq 1$. In practice, this might be impossible to do unless the time series is generated by some simple well-understood mechanism. Instead one typically specifies the first- and second-order moments of the joint distibutions, that is, $\text{E}[X_t]$ and $\text{E}[X_t X_{t+h}]$ for all $t \geq 1$ and for all $h \geq 0$.

The next important concepts are those of stationarity and the auto-correlation function. Roughly speaking, a time series $\{X_t\}_{t=-\infty}^{\infty}$ is stationary if its "statistical properties" are similar to those of the time-shifted series $\{X_{t+h}\}_{t=-\infty}^{\infty}$ for every integer "lag" h. By statistical properties, we mean the first- and second-order moments of $\{X_t\}$.

Formally, a time series $\{X_t\}$ is weakly stationary if

1. $\text{E}[X_t]$ is independent of t
2. $\text{Cov}(X_t, X_{t+h})$ is independent of t for every fixed lag h.

In contrast, a time series $\{X_t\}$ is strictly stationary if the random vectors $(X_{t_1}, \ldots, X_{t_n})'$ and $(X_{t_1+h}, \ldots, X_{t_n+h})'$ have the same joint distributions for all sets of indices $\{t_1, \ldots, t_n\}$, for all $h \geq 0$ and all $n \geq 1$. This is written as:

$$(X_{t_1}, \ldots, X_{t_n})' \stackrel{d}{=} (X_{t_1+h}, \ldots, X_{t_n+h})'$$

Strict stationarity implies the following:

1. The random variables X_t are identically distributed.

2. Pairs of random variables (X_t, X_{t+h}) have the same distribution as (X_1, X_{1+h}) (set $n = 2$).

3. Strict stationarity implies weak stationarity: $\mathrm{E}[X_t] = \mathrm{E}[X_1]$ and $\mathrm{Cov}(X_t, X_{t+h}) = \mathrm{Cov}(X_1, X_{1+h})$ for all $t \geq 1$ and all $h \geq 0$. Both terms are independent of t.

4. Weak stationarity does *not* imply strong stationarity. We show this by an example. Let $Z_i \overset{\text{iid}}{\sim} N(0,1)$ for all i. Define X_t as:

$$X_t = \begin{cases} Z_t & \text{if } t \text{ is even} \\ 2Z_t & \text{if } t \text{ is odd.} \end{cases}$$

   Then $\mathrm{E}[X_t] = 0$ for all t and $\mathrm{Cov}(X_t, X_{t+h}) = 0$, since X_t and X_{t+h} are independent. However, X_0 and X_1 do not have the same distribution.

The autocovariance function of a stationary time series $\{X_t\}$ at lag h is defined as $\mathrm{Cov}(X_{t+h}, X_t) = \mathrm{Cov}(X_h, X_0)$. The autocorrelation of $\{X_t\}$ at lag h is defined as

$$\frac{\mathrm{Cov}(X_{t+h}, X_t)}{\mathrm{Var}(X_t)} = \frac{\mathrm{Cov}(X_h, X_0)}{\mathrm{Var}(X_0)}.$$

3 Bayesian Statistics and MCMC

This section is based on Chapters 12–15 from [4]. Bayes' rule gives us a recipe for calculating the posterior probability density.

$$\mathrm{P}(\Theta \mid \text{data}) = \frac{\mathrm{P}(\text{data} \mid \Theta) \cdot \mathrm{P}(\Theta)}{\mathrm{P}(\text{data})}. \tag{2}$$

Consider a case in which we have a sample of N data points $x_1, \ldots, x_N$. We assume that the likelihood is a Poisson distribution with mean λ and that the prior for λ is a log-normal$(1, 1)$ distribution. To calculate the probability of the data P(data), we must evaluate the integral:

$$\mathrm{P}(\text{data}) = \int_0^\infty \prod_{i=1}^{N} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \cdot \frac{1}{\sqrt{2\pi}\lambda} e^{-\frac{1}{2}(\log\lambda - 1)^2} d\lambda. \tag{3}$$

While this integral is not too difficult, it explains the problem of calculating posteriors analytically. As the number of parameters (the length of Θ) increases, calculating the probability of the data requires evaluating integrals in higher dimensional spaces. This is why we use alternative methods to derive approximate versions of the posterior.

4 Trees, Boosting and Random Forests

5 Neural Networks

References

[1] Peter J. Brockwell, Richard A. Davis. *Introduction to Time Series and Forecasting*, Third Edition, Springer, 2016.

[2] George Casella, Roger L. Berger. *Statistical Inference*, Second Edition, Duxbury Advanced Series, 2001.

[3] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*, Second Edition. Online book at: `https://otexts.com/fpp2/`

[4] Ben Lambert. *A Student's Guide to Bayesian Statistics*, Sage Publications, 2018.