Principle of Data Science

Xinyu Zhong
Queens' College

November 26, 2023

Contents

Abstract

Abstract of this course

Abstract

Abstract of this course

1 Understanding Data

1.1 Visualising Data

1.1.1 Histograms

What are bins in Histogram

1.1.2 Scatter Plots

1.2 Measuring the moments

1.2.1 Average

1.2.2 Spread

1.2.3 Higher Moments

1.3 Covariance and Correlation

Covariance:

$$V_{xy} = cov(x, y) = \overline{xy} - \bar{x}\bar{y} = \frac{1}{N}\sum_{i}^{N}(x_i - \bar{x})(y_i - \bar{y}) \tag{1}$$

Highly covariance means that one variance tends high when the other high, whereas a highly negative covariance means that one variable tends low while the other tends high. A low magnitude of absolute covariance indicates that two variable are more independent of each other.

Covariance Matrix:

$$\boldsymbol{V} = E\left[(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^{\mathrm{T}}\right] = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{1k}\sigma_1\sigma_n & \dots & \dots & \sigma_n^2 \end{bmatrix} \tag{2}$$

Correlation:

$$\rho(x, y) = \frac{\operatorname{cov}(x, y)}{\sigma_x\sigma_y} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x\sigma_y} \tag{3}$$

Correlation is just covariance normalised by the standard deviation product. It is also the sqaure root of R^2

1.4 Learning from Data

1.4.1 Typical Structure of Data

Common data structure include NumPy array or Pandas dataframes.
Note that we have *features* and *events* for a set of data. In pandas, the features correspond to columns whereas event correspond to rows.

1.4.2 Exploiting correlation in data

See Machine Learning for more information.

- Recall: true positive rate. Sometimes also called signal efficiency or sensitivity. The fraction of all positive or signal events that are correctly classified:

$$TPR = \frac{TP}{TP + FN}.$$

- Specificity: true negative rate. Sometimes also called background efficiency. The fraction of all negative or background events that are correctly classified:

$$TNR = \frac{TN}{TN + FP}.$$

- False positive rate. The fraction of all negative or background events that incorrectly classified:

$$FPR = \frac{FP}{FP + TN}.$$

- False negative rate. The fraction of all positive or signal events that incorrectly classified:

$$FNR = \frac{FN}{FN + TP}$$

- Note the relationships between TPR and FNR, as well as TNR and FPR:

$$TPR + FNR = \frac{TP}{TP + FN} + \frac{FN}{FN + TP} = \frac{TP + FN}{TP + FN} = 1,$$
$$TNR + FPR = \frac{TN}{TN + FP} + \frac{FN}{FP + TN} = \frac{TN + FP}{TN + FP} = 1,$$

so that $FPR = 1 - TNR, FNR = 1 - TPR$.

1.4.3 Performance criteria and metrics

- True positive (TP). Correct prediction of positive or signal outcome.
- False positive (FP). Incorrect prediction of positive or signal outcome.
- True negative (TN). Correct prediction of negative or background outcome.
- False negative (FN). Incorrect prediction of negative or background outcome.
- All positive or signal events are given by $P = TP + FN$.
- All negative or background events are given by $N = TN + FP$.
- All events classified as positive or signal-like are given by $C_P = TP + FP$.
- All events classified as negative or signal-like are given by $C_N = TN + FN$.
- Recall(or signal efficiency in particle physics): true positive rate. Sometimes also called signal efficiency or sensitivity. The fraction of all positive or signal events that are correctly classified:

$$TPR = \frac{TP}{TP + FN}.$$

- Specificity(or): true negative rate. Sometimes also called background efficiency. The fraction of all negative or background events that are correctly classified:

$$TNR = \frac{TN}{TN + FP}.$$

- False positive rate. The fraction of all negative or background events that incorrectly classified:

$$FPR = \frac{FP}{FP + TN}.$$

- False negative rate. The fraction of all positive or signal events that incorrectly classified:

$$FNR = \frac{FN}{FN + TP}$$

- Note the relationships between TPR and FNR, as well as TNR and FPR:

$$TPR + FNR = \frac{TP}{TP + FN} + \frac{FN}{FN + TP} = \frac{TP + FN}{TP + FN} = 1,$$
$$TNR + FPR = \frac{TN}{TN + FP} + \frac{FN}{FP + TN} = \frac{TN + FP}{TN + FP} = 1,$$

- Accuracy. The fraction of all events that correctly classified:

$$\alpha = \frac{TP + TN}{P + N}.$$

- Error rate. The fraction of all events that are incorrectly classified:

$$\varepsilon = \frac{FP + FN}{P + N}.$$

- Purity. The fraction of all events classified positively that are correctly classified:

$$\rho_P = \frac{TP}{TP + FP} \quad \text{and} \quad \rho_N = \frac{TN}{TN + FN}.$$

- Significance. For a counting experiment quantifies the statistical significance:

$$\sigma = \frac{TP}{\sqrt{TP + FP}}.$$

- Signal-to-noise, sometimes also called signal-to-background ratio:

$$SNR = \frac{TP}{FP}.$$

- F-score, sometimes also called F-measure:

$$F = 2\frac{\text{precision } \times \text{ recall}}{\text{precisoin } + \text{ recall}} = \frac{2TPR}{2TPR + FPR + FNR}.$$

Concepts of Type I and Type II errors.
Usually, one algorithm will prioritise one thing at the cost of the other, for example maximise TPR will results in a high FPR as well.
A good metric to optimise is ROC (receiver-operating-characteristic), which is the curve in FPR vs TPF graph, and we would like to push the curve to the top left.

1.4.4 Data challeneges

2 Probability

2.1 Definition of Probability

2.1.1 Frequentist

Frequentist - The true probability in practise is never obtainable (because we cannot perform infinite experiments). We can only estimate the probability given the sample size we have. However, if it always possible to perform more experiments then we can keep doing so until we reach the desired accuracy (and any accuracy is in principle achievable). - Frequentist probability can only be applied to repeatable experiments. For example I cannot use it to determine the probability it will rain the day after tomorrow. I need a system in which I can keep relevent conditions stable to perform repeatable experiments. - One does not need to have any prior beliefs about outcomes, the probability is purely determined from observations.

2.1.2

Baysian Finetti's coherent bet:

- 1. If you are willing to bet on the outcome of a random experiment, then you should be willing to bet on the outcome of any exchangeable random experiment.

2.2 Property of Probability

Property of probability is based on Kolmogorov's axioms

- Addition
- Conditional Probability
- Independence A and B are independent if P(A, B) = P(A)P(B)

2.2.1 Monty Hall Problem

Check Example Sheet 1

2.3 Probability mass/density Function

$P(X)$ to denote probability mass function (discrete probability) $p(X)$ to denote probability density function (continuous probability)

2.4 Change of variables

$$X \ f(x)$$

Denotes that X is distributed like f(x)

$$y = h(x)$$

Denotes that y is a function of X To change variables, probability must be conserved Functions are invertible

** Jacobian Matrix

2.4.1 The cumulative distribution

Cumulative distribution function:

$$F(X) = \int_{-\infty}^{X} f(X')\, dX' \tag{4}$$

which is the integrated p.d.f.

It is defined so that

$$F(X_{\min}) = 0,$$
$$F(X_{\max}) = 1,$$

with some very useful properties

$$P(X < X') = \int_{X_{\min}}^{X'} f(X)dX = F(X'),$$
$$P(X' < X < X'') = \int_{X'}^{X''} f(X)dX = F(X'') - F(X).$$

The integral of p.d.f is the difference in c.d.f.

2.4.2 The Joint, Marginal and conditional distribution

independent indicates uncorrelated ? not correlated does not indicate independence (coreelated means linearly-independent, they can have a quadratic relationship)

Correlation means linear relationships
Joint Probability:

$$f(X, Y) = g(X)h(Y) \tag{5}$$

Marginal Distribution:

$$g(X) = \int f(X, Y)dY \text{ the marginal distribution in } X \tag{6}$$
$$h(Y) = \int f(X, Y)dX \text{ the marginal distribution in } Y. \tag{7}$$

Conditional distribution:

$$g(X \mid Y) = \frac{f(X, Y)}{h(Y)} = \frac{f(X, Y)}{\int f(X, Y)dX} the probability of X given Y, \tag{8}$$
$$h(Y \mid X) = \frac{f(X, Y)}{g(X)} = \frac{f(X, Y)}{\int f(X, Y)dY} the probability of Y given X. \tag{9}$$

2.4.3 Bayes' theorem for continuous variables

Bayes' theorem:

$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{p(X)} = \frac{p(X \mid \theta)p(\theta)}{\int p(X \mid \theta)p(\theta)d\theta} \tag{10}$$

The terms in the equation

- $p(\theta \mid X)$ - the *posterior distribution* - our probability distribution for the parameter θ given the data we have observed X. 50
- $p(X \mid \theta)$ - the *likelihood function* - the likelihood we observe the data X given a particular value of θ. This is of vital importance across all of statistics and machine learning. We will discuss the likelihood in much more detail in Sec. 3.2.
- $p(\theta)$ - the *prior distribution* - encompassing our prior beliefs about θ, this can be based on previous measurements, previous beliefs or indeed be flat (although note that a change of variables or basis will not necessarily maintain flatness). The prior influences the outcome of Bayesian inference, which can be seen as an advatange and a disadvantage. I will leave discussion of priors to the other stats course.
- $p(X)$ - the *evidence* - this is just a normalisation factor that ensures the posterior is a p.d.f.. For Bayesian inference the evidence can often be ignored as the posterior is proportional to the numerator of (2.40). However, the evidence can be quite a useful quantity for goodness of fit tests. I will leave discussion of the evidence to the other stats course.

.

2.5 Properties of Distributions

2.5.1 Expextation, mean adn variance

Expectation:

$$E[g(X)] = \int g(X)f(X)dX \tag{11}$$

Expertation is a linear operator, is the true mean of the distribution

Variance:

$$\begin{aligned} V(x) = \sigma^2 &= E\left[(X-\mu)^2\right] && (12)\\ &= E\left[X^2 - 2\mu X + \mu^2\right] && (13)\\ &= E\left[X^2\right] - \mu^2 && (14)\\ &= \sigma^2 && (15) \end{aligned}$$

We can also define the moments of the distribution using expectation values. The moments are defined as

$$\begin{array}{ll} \mu_\ell = E\left[X^\ell\right] & \text{the } \ell^{\text{th}} \text{ algebraic moment} \\ \alpha_\ell = E\left[(X - E[X])^\ell\right] & \text{the } \ell^{\text{th}} \text{ moment about the mean.} \end{array}$$

In our notation the mean is $\mu = \mu_1$ and the variance is $\sigma^2 = \alpha_2$. The skew is $\gamma_1 = \sqrt{\beta_1} = \mu_3/\mu_2^{3/2}$. The kurtosis is $\gamma_2 = \beta_2 - 3 = \mu_4/\mu_2^2 - 3$.

2.5.2 Covariance and Correlation

We can also define the covariance between two random variables X and Y,

$$V(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right] = E[XY] - E[X]E[Y],$$

and then the correlation

$$\rho(X,Y) = \frac{V(X,Y)}{\sigma_X \sigma_Y}$$

Note the math behind derivation.

Example Assume $x \sim f(x) = Ne^{-x^2}$
Find $N \Rightarrow$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \Rightarrow N \Rightarrow \frac{1}{\sqrt{\pi}}$$

Find $E[x] \Rightarrow$

$$\int_{-\infty}^{\infty} xf(x)dx = 0$$

as its add function.
Find $E\left[(x-\mu)^2\right] \Rightarrow$

$$E\left[x^2\right] - (E[x])^2$$

$$E\left[x^2\right] = \int_{-\infty}^{\infty} x^2 f(x)dx \quad \Rightarrow \quad \sigma^2 = \frac{1}{2}$$

To get $\sigma = 1$. $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \rightarrow$ stanclond model shift x by μ and scale by $\frac{1}{\sigma}$. ie. $z = \frac{x-\mu}{\sigma} \Rightarrow$ Now Model

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\frac{1}{N} = \sigma\sqrt{2\pi}, \text{ Mean } = \mu \quad \text{variance } = \sigma^2$$

2.5.3 The characteristic function

characteristic function:

$$\varphi(t) = E\left[e^{itX}\right] = \int_{-\infty}^{\infty} e^{itX} f(X)dX \tag{16}$$

which means that $f(X)$ is completely defined by the characteristic functions

$$f(X) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi(t) e^{-iXt} dt$$

The usefulness of the characteristic function is shown in proof for central limit theorem, or algebraic moments

$$\mu_n = E\left[X^n\right] = \int_{-\infty}^{\infty} X^n f(X)dX$$

which can be obtained by differentiating the characteristic function n times at point $t = 0$

$$\varphi_n(t) = \frac{d^n \varphi(t)}{dt^n} = i^n \int_{-\infty}^{\infty} x^n e^{itX} f(X)dX$$

such that $\varphi_n(0) = i^n \mu_n$.

2.6 Common Distribution

p.d.f.s depend on one r.v.s. x and parameters θ, write as

$$p(x;\theta)$$

where ; distinguish rvs and parameters

2.6.1 Binomial Distribution

Binomial distribution:

$$P(k;p,n) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \tag{17}$$

given n trials,$p(sucess) = p$, $p(fail) = q = 1-p$ and the total probability of k triads are success is

$$p^k(1-p)^{n-k}$$

2.6.2 Poisson Distribution

Poisson distribution:

$$P(k,\lambda) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{18}$$

Derivation of Poisson Distribution from binomial

$$P(k;\lambda/n,n) = \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \frac{n!}{k!(n-k)!}$$

Mean of Poisson Distribution : λ
Variance of Poisson Distribution : λ

2.6.3 Normal Distribution

Normal Distribution:

(19)

It can be shifted by μ and scale by σ

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$f(x) = \phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-z^2/2}dz$$

p-value is the probability values

2.6.4 Multi-variate Normal Distribution

for independent r.v.s. $x_1, x_2, \ldots, x_n$

$$p(\vec{\boldsymbol{x}};\vec{\boldsymbol{\mu}},\vec{\boldsymbol{\sigma}}) = \prod_{i=1}^{n} p(x_i;\mu_i,\sigma_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}}$$

with dependent r.v.s. have a correlation term: $\vec{\boldsymbol{\sigma}} =$

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix}$$

Terms in exp becomes $(x-\mu)^T V^{-1}(x-\mu)$

2.6.5 The exponential decay Distributions

2.6.6 Polynomial distributions

2.6.7 Chi-squared Distribution

Chi-squared Distribution:

$$P(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} \tag{20}$$

Chi-squared distribution with k degree of freedom gives the distribution of the sum of squares of k independent standard normal variables

2.6.8 Generating samples from distributions and the inverse c.d.f.

using inverse c.d.f. to generate samples from distributions or use accept and reject method

2.7 Limit Theorems

2.7.1 Convergence

2.7.2 Central Limit Theorem

2.7.3 Errors

3 Classical Statistics, Estimation and Uncertainties

3.1 Frequentist evaluation of estimators

3.1.1 Consistency, bias and efficiency of estimates

Consistency A Consistent estimator will converge to true value as datasize increases:

$$\lim_{N\to\infty} \hat{\theta}(\vec{x}) = \theta$$

Bias The bias of an estimator is the difference between the expected value of the estimator and the true value of the parameter being estimated:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

It is generally preferrable to have consistent estimator as opposed to biased estimators.

Efficiency The efficiency of an estimator is the ratio of the variances of two unbiased estimators. The estimator with the lower variance is said to be more efficient.

$$\text{Eff}(\hat{\theta}) = \frac{\text{Var}(\hat{\theta})}{\text{Var}(\hat{\theta}')}$$

3.1.2 Bias-Variance Trade-off

In order to minimise the variance, we want to take in more data, however, the more data we take in, the more bias we introduce. This is the bias-variance trade-off.

3.1.3 Estimation of the mean, variance and standard deviation

3.2 The likelihood function

The likelihood function:

$$p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta) \tag{21}$$

The likelyhood function is a function of θ only, and is the probability of observing the data given the parameter θ

3.3 Minimym Variance Bound

3.3.1 Maximum Likelihood Estimation

Maximise L is equivalent to maximise $\ln(L)$

3.4 Least-square method

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(x_i))^2}{\sigma_i^2}$$

σ_{y_i} is the associated uncertainty of y_i. We will see that it is liked to the log likelihood function by

$$\chi^2 = -2 \ln L + C$$

note that both side of the equation is a function of θ (estimated ?) only.
This allows us to obtain the relationship in difference:

$$\Delta\chi^2 = -2\Delta \ln L$$

Wilks' theorem:

As $N \to \infty$, the test-statistic which is twice the negative log likelihood ratio approached the χ^2 distribtion. This is an example of hypothesis test where the null hypothesis is the valule of the parameter at the best fit and the alternative hypothesis is the value of the parameter where we read off the log likelihood difference.

3.4.1 Fisher information

Fisher information:

$$I(\theta) = E\left[\left(\frac{\partial \ln(p(X;\theta))}{\partial\theta}\right)^2\right] \tag{22}$$

The Fisher information is the expected value of the squared score function, which is the derivative of the log likelihood function.

3.5 Method of Moments

Method of moments is a method of estimating the parameters of a statistical model. Foe example, give a distribution of known form but unknown parameters,

$$f(x;\vec{\theta})$$

We can estimate the parameters by equating the sample moments to the theoretical moments.

$$1 = \int_{-\infty}^{\infty} f(x;\vec{\theta})dx \tag{23}$$

$$\mu = \int_{-\infty}^{\infty} xf(x;\vec{\theta})dx = \text{some funciton of theta}, i.e. g(\mu,\sigma) \approx \hat{\mu} \tag{24}$$

$$\mu^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x;\vec{\theta})d = \text{some function of theta} \approx \hat{\mu}^2 \tag{25}$$

$$... \tag{26}$$

Here $\vec{\theta}$ is a vector of parameters, hence we need the right amount (number of independent parameters, i.e. μ,σ) of moments to estimate the parameters. We can estimate the parameters by equating the sample moments $\hat{\mu}$ to real moments μ, and solve simultaneous equations to get the parameters.

3.5.1 Uncetainties from method of moments estimates

3.6 Goodness-of-Fit Tests

This section will discuss the question of "how good was my fit in the first place", the answer to which is the goodness-of-fit test.

Test Statistics:

$$\text{Quantify the agreement between the data and the model} \tag{27}$$

i.e. from the probability distribution, we can compute the probabilitay that we got the value we did, which in term gives us info on what we think the quality of the fit. Example of Test Statistics include χ^2

3.6.1 χ^2 Tests

χ^2 tests is chosen as we expect $\chi^2/d.o.f = 1$, as the expectation of the χ^2 for k degree of freedom is k.

The χ^2 probability, i.e. $1-F(\chi^2)$ (1 - c.d.f. of the χ^2 distribution of the appropriate degree of freedom). What the value p means is the probability that a function that does describe the data gives a value of the χ^2 larger than the one you find (A larger χ^2 means you did a bad job fitting).

For p-value:

- if p-value is small, the model do not agree well
- if p-value is large, the model and data agree too well, over fitting the data.

χ^2 test is very good at spotting the small p-value.

This test is very dependent on the binning that us used and it is bad if you are trying to assess the compatibility of a model with may only contain a discrepancy in one bin. Example in Higgs Boson

3.6.2 Kolmogorov-Smirnoff KS Tests

It is an unbinned test, so it can be an alternative to the χ^2 test when the number of the event is small. Fins the maximum difference between the two cumulative distributions. *KS score:*

$$P_{\text{KS}} = \max |F_1(x) - F_2(x)| \sqrt{N} \tag{28}$$

which is the maximum deviation between the two distributions cdf and multiplying bu the square root of the sample size.

3.7 Confidence Intervals

Whenever we qant to quote a point estimate, we should also quote an interval estimate. e.g. The common practise is quote 68.3% or 95.4% confidence interval.

3.7.1 Bayesian Intervals

3.7.2 Classical Intervals (Neyman-Pearson intervals)

Classical intervals are based on the Neyman-Pearson lemma, which states that the likelihood ratio test is the most powerful test for a given significance level α.

3.7.3 Feldman-Cousins intervals

3.8 Hypothesis Testing

3.8.1 Neyman-Pearson Lemma

Neyman-Pearson Lemma:

The likelihood ratio test is the most poweful test

The Neyman-Pearson Lemma addresses the case where you want to construct a test with the maximum power subject to a specified Type I error rate (α). In other words, it helps you design the most powerful test for detecting a particular effect, given a constraint on the probability of making a Type I error.

3.9 Resampling Methods

Resampling is the process of drawing data from a sample to produce additional samples.

4 Advanced Topics

4.1 Measurement errors and forward modelling

4.2 Optimisation

4.3 Regularisation

4.4 Density Estimation

This chapter uses the Kernal Density Estimation (KDE) method to estimate the probability density function of a random variable,which replaces the step function.

4.4.1 Kernal Density Estimation

Kernal Density Estimation:

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{x-x_i}{h}\right) \tag{29}$$

where K is the kernal function, h is the bandwidth, and N is the number of data points.