Machine Learning Methods

Siger

September 1, 2023

Si Dieu est infini, alors je suis une partie de Dieu sinon je serai sa limite...

Contents

Chapter 1

Collect and Pre-process Data

1.1 Data cleaning

[1]

1.1.1 Data Quality

Validity

- **Data-Type Constraints**: for a given column a fixed data-type must be associated with.
- **Range Constraints**: only a range of values should be taken.
- **Mandatory Constraints**: some columns cannot be empty.
- **Unique Constraints**: across a given dataset a field or a combination of
- **Foreign-key constraints**: a foreign key column cannot have a value that does not exist in the primary key.
- **Regular expression patterns**: text fields that have to follow a given alphanumerical pattern.
- **Cross-field validation**: consistency of values, for example considering a given man, his birth date have to be older than his death date.

Accuracy The degree to which the data is close to the true value.

Completeness The degree to which the all the required data is known.

Consistency The degree to which the data is consistent, within the same data set or across multiple data sets.

Uniformity The degree to which the data is specified using the same unit of measure.

1.1.2 The workflow

Inspection Detect unexpected behavior in the data.

- **Data profiling**: summary statistics about the data, see ydata-profiling in Python.
- **Visualizations**: visualize the data using statistical metrics, see plotly
- **Software packages**: to note and check the constraints regarding the data see pydeequ

Cleaning Fix or remove anomalies discovered in the above phase.

- **Irrelevant Data**: ask to the expert what can be the unnecessary columns, check them and remove them if they are not useful.
- **Duplicates**
- **Type conversion**: make sure the appropriate data type is associated with a given column.
- **Syntax errors**: white spaces, pad strings ...
- **Standardize**: same unit across the dataset, same pattern for text.
- **Scaling/Transformation**: in order to compare different scores for example.
- **Normalization**: useful for some statistical methods.
- **Missing values**:
  - Drop: only if the missing values in a column rarely and randomly occur.
  - Impute: many methods, *mean* is relevant when data is not skewed otherwise we should use *median*. A linear regression or a hot-deck (copying of values) approach can be taken as well, and more interestingly a *k-nearest* method approach.
  - Flag: let the missing value as it is.
- Outliers: Remove outliers only if they are harmful for the chosen model.
- In-record & cross-datasets errors: fix non-consistent situations like married kids, quantity being different of the one when we compute using other columns.

Verifying Check correctness of the cleaning phase.

Reporting Report about changes made, using one of the software summarising the data quality for example.

Chapter 2

Statistics

2.1 Fundamental probability concepts

2.1.1 Basic probability properties

What is a probability? It is a mathematical measure of the uncertainty of a given event.

Objectivist interpretation [3] : assigns numbers describing some objective state, *Frequentist* interpretation claiming that the probability of a random event is quantified by the relative frequency in a given experiment.

Subjectivist interpretation [3] : assigns numbers quantifying the degree of belief that a given event occurs. *Bayesian* interpretation uses expert knowledge considered as subjective and represented by the prior, as well as experimental data represented by the likelihood. The normalized product of the 2 above quantity is the posterior probability distribution containing both expert knowledge and experimental data.

Properties

Event and its opposite $\mathbb{P}(\{A\}) + \mathbb{P}\left(\{\overline{A}\}\right) = 1$

Not necessary mutually exclusive events $\mathbb{P}(\{A \cup B\}) = \mathbb{P}(\{A\}) + \mathbb{P}(\{B\}) - \mathbb{P}(\{A \cap B\})$

Independent events $\mathbb{P}(\{A \cap B\}) = \mathbb{P}(\{A\}) \times \mathbb{P}(\{B\})$

Conditional Probability $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

Law of Total Probability $\begin{cases} (B_i)_{1 \leq i \leq n} : \text{ partition of a sample } \mathcal{S} \\ \forall i \in [\![1, n]\!],\ \mathbb{P}(\{B_i\}) \neq 0 \end{cases} \Rightarrow \mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(B_i)\,\mathbb{P}(A|B_i)$

Bayes' Theorem Using Law of Total Probability:

$$\begin{cases} (B_i)_{1 \leq i \leq n} : \text{ partition of a sample } \mathcal{S} \\ \forall i \in [\![1, n]\!],\ \mathbb{P}(\{B_i\}) \neq 0 \end{cases} \Rightarrow \mathbb{P}(B_i|A) = \frac{\mathbb{P}(B_i) \times \mathbb{P}(A|B_i)}{\sum_{k=1}^{n} \mathbb{P}(B_k)\,\mathbb{P}(A|B_k)}$$

Moments They are certain quantitative measures related to the shape of the function's graph. [2]

n^{th} moments of a random variable: The n^{th} moment about the origin of a random variable X as denoted by $E(X^n)$, is defined to be:

$$\mathbb{E}(X^n) = \begin{cases} \sum_{x \in R_X} x^n f(x) \text{ if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^n f(x) dx \text{ if } X \text{ is continuous} \end{cases}$$

Expected value: The expected value of a random variable X as denoted by $E(X)$, is defined to be:

$$\mathbb{E}(X) = \begin{cases} \sum_{x \in R_X} x f(x) \text{ if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f(x) dx \text{ if } X \text{ is continuous} \end{cases}$$

After normalized this moment by total mass we have the center of mass.

Variance : Let X be a random variable with mean μ_X. The variance of X denoted by $\mathbb{V}(X)$ or σ_X^2 is defined by:

$$\mathbb{V}(X) = \mathbb{E}\left([X - \mu_X]^2\right)$$

After normalized this moment by total mass we have the moment of inertia.
If X is a random variable with mean μ_X and variance σ_X^2 then:

$$\sigma_X^2 = \mathbb{E}(X^2) - \mu_X^2$$

And:

$$\mathbb{V}(aX + b) = a^2 \mathbb{V}(X)$$

Skewness and Kurtosis

- Skewness: $\mathbb{E}\left(\left[\frac{X - \mu_X}{\sigma_X}\right]^3\right)$, indicates the direction (negative → left tail is longer, positive → right tail is longer) and relative magnitude of a distribution's deviation from the normal distribution.
- Kurtosis: $\mathbb{E}\left(\left[\frac{X - \mu_X}{\sigma_X}\right]^4\right)$, measures the outliers, data within one standard deviation will not contribute a lot to the kurtosis values conversely data exceeding one standard deviation will contribute a lot because of the fourth power.

Asymptotic properties

Chebychev inequality allows to find an estimate of the area between the values $\mu - k\sigma$ and $\mu + k\sigma$ for some given $k \neq 0$, showing that the area under $f(x)$ on the interval $[\mu - k\sigma, \mu + k\sigma]$ is at least $1 - k^2$. Let X be a random variable with probability density function $f(x)$. If μ and $\sigma > 0$ are the mean and standard deviation of X then:

$$\boxed{\mathbb{P}(\{|X - \mu| < k\sigma\}) \geq 1 - \frac{1}{k^2}}$$

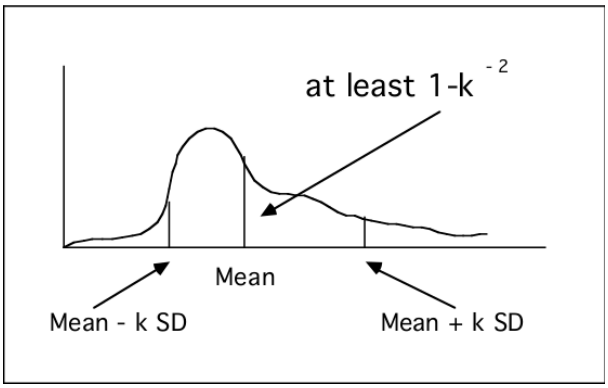

Figure 2.1: Illustration of Chebychev inequality

Markov inequality

$$X \neq \underline{0} \Rightarrow \mathbb{P}\left(\{X \geq t\}\right) \leq \frac{\mathbb{E}\left(X\right)}{t}$$

Theorem weak law of large numbers: Let $(X_i)_{1\leq i\leq 1}\, n$: independent & identically distributed RV

$$\lim_{n\to\infty} \mathbb{P}\left(\left\{\left|\overline{S}_n - \mu\right| \geq \epsilon\right\}\right) = 0 \text{ with } \overline{S}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Convergence in probability Suppose $(X_i)_{1\leq i\leq 1}\, n$ is a sequence of random variables de- fined on a sample space S. The sequence "converges in probability" to the random variable X if, for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(\{|X_n - X| < \epsilon\}\right) = 1$$

Convergence almost surely Suppose the RV X and $(X_i)_{1\leq i\leq 1}\, n$ is a sequence of random variables de- fined on a sample space S. The sequence $X_n(\omega)$ "converges almost surely" to $X(\omega)$ if

$$\mathbb{P}\left(\left\{w \in S | \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1$$

Properties

- For a Bernoulli distribution, $\overline{S}_n$ converges in probability to p
- For a Normal distribution, $\overline{S}_n$ converges almost surely to μ

Central Limit Theorem The central limit therorem (Lindeberg-Levy Theorem) states that for any population distribution, the distribution of the standardized sample mean is approximately standard normal with better approximations obtained with the larger sample size.

$$\begin{cases} (X_i)_{1\leq i\leq 1}\, n \hookrightarrow ?(\mu, \sigma^2) \\ n \to \infty \end{cases} \quad \Rightarrow \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \hookrightarrow \mathcal{N}(0,1)$$

Convergence in distribution Consider X with its cumulative density function F and $(X_i)_{1\leq i\leq 1}\, n$ with their cdf $(F_i)_{1\leq i\leq n}$:

$$\lim_{n\to\infty} F_n(x) = F(x) \Rightarrow X_n \text{ "converges in distribution" to } X$$

Lévy Continuity Theorem

$$\begin{cases} (X_i)_{1\leq i\leq 1}\, n\text{RV} \\ (F_i)_{1\leq i\leq 1}\, n\text{distribution functions} \\ (M_{X_i})_{1\leq i\leq n}\, \text{moment generating function} \end{cases} \qquad \forall t \in [-h,h] \lim_{n\to\infty} M_{X_n}(t) = M_X(t) \Rightarrow \lim_{n\to\infty} F_n(x) = F(x)$$

Bivariate case

Joint probability density function Let $(X,Y):(\Omega_X,\Omega_Y)\to(R_X,R_Y)$ and $f:R_X\times R_Y\to\mathbb{R}$

$$\forall(x,y)\in R_X\times R_Y, f(x,y)=\mathbb{P}\left(\{X=x,Y=y\}\right)\Leftrightarrow$$
f is the joint probability density function for X and Y

Marginal probability density function Let for all $(x,y)\in R_X\times R_Y$: $f(x,y)$ be the joint probability density of X and Y

$$\begin{cases} f_1(x)=\displaystyle\int_{-\infty}^{\infty} f(x,y)dy \text{is the marginal probability density of } X \\ f_2(y)=\displaystyle\int_{-\infty}^{\infty} f(x,y)dx \text{is the marginal probability density of } Y \end{cases}$$

Joint cumulative probability distribution function Let $F:\mathbb{R}^2\to\mathbb{R}$

$$\forall(x,y)\in\mathbb{R}^2, F(x,y)=\mathbb{P}\left(\{X\leq x,Y\leq y\}\right)=\int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v)dudv\Leftrightarrow$$
F is the joint cumulative probability density function for X and Y

From the fundamental theorem of calculus: $f(x,y)=\dfrac{\partial^2 F(x,y)}{\partial x\partial y}$

Conditional expectation The conditional mean of X given $Y=y$ is defined as:

$$\mathbb{E}\left(X|y\right)=\begin{cases}\displaystyle\sum_{x\in R_X} xg(x/y)\Leftarrow X \text{ discrete}\\ \displaystyle\int_{-\infty}^{\infty} xg(x/y)dx\Leftarrow X \text{ continuous}\end{cases}$$

Properties:

$$\begin{cases}\mathbb{E}_X\left(\mathbb{E}_{y|x}\left(Y|X\right)\right)=\mathbb{E}_y\left(Y\right)\\ \mathbb{E}\left(Y|\left\{X=x\right\}\right)=\mu_Y+\rho\dfrac{\sigma_Y}{\sigma_X}(x-\mu_X)\end{cases}$$

Conditional Variance

$$\begin{cases}\mathbb{V}\left(Y|x\right)=\mathbb{E}\left(Y^2|x\right)-\mathbb{E}\left(Y|x\right)^2\\ \mathbb{E}_x\left(\mathbb{V}\left(Y|X\right)=(1-\rho^2)\mathbb{V}\left(Y\right)\right)\end{cases}$$

2.1.2 Distribution function

Definition of probability density function (pdf): Let R_X be the space of the random variable X. The function: $f:R_X\to\mathbb{R}$ defined by:

$$f(x)=\mathbb{P}\left(\{X=x\}\right) \text{ if } X \text{ is discrete.}$$
$$f(x)=\mathbb{P}\left(\{X\in A\}\right)=\int_A f(x)dx \text{ if } X \text{ is continuous, with } A \text{ a set of real numbers.}$$

is called probability density function of X.

Definition of cumulative density function (cdf): Let R_X be the space of the random variable X. The function: $F:R_X\to\mathbb{R}$ defined by:

$$F(x)=\mathbb{P}\left(\{X\leq x\}\right) \text{ if } X \text{ is discrete.}$$
$$F(x)=\mathbb{P}\left(\{X\leq x\}\right)=\int_{-\infty}^{x} f(t)dt \text{ if } X \text{ is continuous, with } A \text{ a set of real numbers.}$$

Percentile for continuous random variables. Let $p \in [0;1]$, a $100p^{th}$ percentile of the distribution of a random variable X is $q \in \mathbb{R}$ satisfying:

$$\mathbb{P}(\{X \leq q\}) \leq p$$

(Recall that the F is a monotonically increasing function, then it has an inverse F^{-1})

$$q = F^{-1}(p)$$

A $100p^{th}$ is a measure of location for the probability distribution in the sense that q divides the distribution of the probability mass into 2 parts, one having probability mass p and other having probability mass $1 - p$

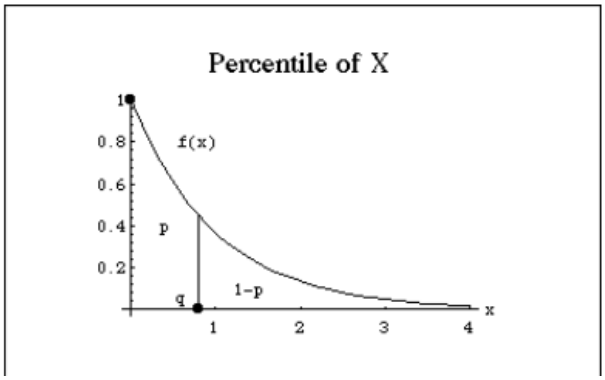

Figure 2.2: Percentile

The 50^{th} percentile of any distribution is called median of the distribution.

2.2 Distributions

2.2.1 Discrete distributions with finite support

Bernoulli

Rademacher

Binomial

Beta-Binomial

Degenerate

Uniform

Hypergeometric

Negative Hypergeometric

Poisson Binomial

Fisher's noncentral hypergeometric

Benford's law

Zipf's law

Zipf-Mandelbrot law

2.3 Bayesian approach

2.3.1 Components

Bayesian concept learning Let be $\mathcal{D}$ the data, h the hypothesis taken in account

Likelihood $p(\mathcal{D}|h)$ the probability to get the observed data considering the hypothesis h.

Prior $p(h)$ the probability of our hypothesis, many prior can be used, and this **subjective** aspect of Bayesian reasoning is a source of much controversy.

Posterior The posterior is simply the likelihood times the prior, normalized.

$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h) \times p(h)}{\sum_{h' \in \mathcal{H}} p(\mathcal{D}, h')\, p(h')}$$

2.3.2 Summarizing posterior distributions

MAP (Maximum A Posteriori) estimation Although most appropriate choice for:

$$\begin{cases} \text{Real valued quantity} & \rightarrow \textit{posterior median or mean} \\ \text{Discrete} & \rightarrow \textit{vector of posterior marginals} \end{cases}$$

The most popular choice is *posterior mode* aka MAP, because it reduces to optimization problems for which efficient algorithms often exist.
Some point to be aware about MAP:

- No measure of uncertainty
- Plugging in the MAP estimate can result in overfitting
- The mode is an untypical point, unlike the mean or median the mode is a point of measure 0, it does not take the volume of the space into account.
- MAP estimation is not invariant to reparameterization, for example passing from centimeters to inches can break things.)
  The MLE does not suffer from this since the likelihood is a function not a probability density

Credible intervals With point estimates, we want a measure of confidence.

$$C_\alpha(\mathcal{D}) = (l, u) : \mathbb{P}(\{l \leq \theta \leq u|\mathcal{D}\})$$

In general, credible intervals are usually what people want to compute but confidence intervals are usually what they actually compute, because most people are taught frequentist statistics but not Bayesian statistics.
Sometimes with central intervals there might be points be outside the CI which have higher probability density.
More formally p^* such that:

$$1 - \alpha = \int_{\theta : p(\theta|\mathcal{D}) > p^*} p(\theta|\mathcal{D}) d\theta$$

Then the HPD such that:

$$\mathcal{D} = \{\theta : p(\theta|\mathcal{D}) \geq p^*\}$$

2.3.3 Bayesian Model Selection

A more efficient approach than cross-validation, meaning fitting k times each model, is to compute the posterior over models.

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m \in \mathcal{M}} p(m|\mathcal{D})}$$

From this we can compute the MAP model $\hat{m} = \arg\max_m p(m|\mathcal{D})$

Then we have the marginal likelihood: $p(\mathcal{D}|\hat{m}) = \int p(\mathcal{D}|\hat{m})p(\theta|\hat{m})d\theta$

Baysian Occam's razor In integrating out the parameters rather than maximizing them we are automatically protected from overfitting: model with more parameters do not necessarily have higher marginal likelihood.
A way to understand the Bayesian Occam's razor effect is to remember that probabilities must sum to one, meaning $\sum_{\mathcal{D}'} p(\mathcal{D}'|m) = 1$. Complex models, which can predict many things, must spread their probability mass thinly, and hence will not obtain as large a probability for any given data set as simpler models.

Computing the marginal likelihood (evidence) For a fixed model we often write:

$$p(\boldsymbol{\theta}|\mathcal{D}, m) \propto p(\boldsymbol{\theta}|m)p(\mathcal{D}|\boldsymbol{\theta}, m)$$

This valid since $p(\mathcal{D}|m)$ is constant. However when comparing models we need to know how to compute the marginal likelihood, $p(\mathcal{D}|m)$. In general this can be quite hard, since we have to integrate over all possible parameter values, but when we have a conjugate prior, it is easy to compute.
Let $p(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{Z_0}$ be our prior, where $q(\boldsymbol{\theta})$ is an unnormalized distribution, and Z_0 is the normalization constant of the prior. Let $p(\mathcal{D}|\boldsymbol{\theta}) = \frac{q(\mathcal{D}|\boldsymbol{\theta})}{Z_l}$ be the likelihood, where Z_l contains any constant factors in the likelihood. Finally let $p(\boldsymbol{\theta}|\mathcal{D}) = \frac{q(\boldsymbol{\theta}|\mathcal{D})}{Z_N}$ be our posterior where $q(\boldsymbol{\theta}|\mathcal{D}) = q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})$ is the unnormalized posterior, and Z_N is the normalization constant of the posterior.

We have: $\begin{cases} p(\boldsymbol{\theta}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \\ \frac{q(\boldsymbol{\theta}|\mathcal{D})}{Z_N} = \frac{q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})}{Z_l Z_0 p(\mathcal{D})} \\ p(\mathcal{D}) = \frac{Z_N}{Z_0 Z_l} \end{cases}$

In general $p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$ can be quite difficult to compute. Simpler approach

- **BIC** simple approximation: $BIC \triangleq \log(p(\mathcal{D}|\hat{\boldsymbol{\theta}})) - \frac{dof(\hat{\boldsymbol{\theta}})}{2}\log(N) \approx \log p(\mathcal{D})$
- **AIC**: $AIC(m, \mathcal{D}) \triangleq \log(p(\mathcal{D})\hat{\boldsymbol{\theta}}_{MLE}) - dof(m)$
  This is derived from Frequentits framework and cannot be interpreted as an approximation to the marginal likelihood. The penalty of AIC is less than BIC, it causes AIC pick more complex models. That can be better for predictive accuracy.
- Effect of the prior.
  If the prior is unknown, the correct Bayesian procedure is to put a prior on the prior. That is we should put a prior on the hyper-parameter α as well as the parameters $\boldsymbol{w}$. To compute the marginal likelihood we should integrate out all unknowns, we should compute: $\int\int p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}|\alpha, m)p(\alpha|m)d\boldsymbol{w}d\alpha$
  A computational shortcut is to optimize α rather than integrating it out. That is, we use $p(\mathcal{D}|m) \approx \int p(\mathcal{D}\boldsymbol{w})p(\boldsymbol{w}|\alpha, m)d\boldsymbol{w}$. where $\hat{\alpha} = \arg\max_\alpha p(\mathcal{D}|\alpha, m) = \arg\max_\alpha \int p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}|\hat{\alpha}, m)d\boldsymbol{w}$

Bayes Factors When prior on models is uniform, then model selection is equivalent to picking the model with the highest marginal likelihood. Now suppose we just have two models we are considering, call them the null hypothesis, M_0 and the alternative hypothesis, M_1.

$$BF_{1,0} \triangleq \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)} = \frac{\frac{p(M_1|\mathcal{D})}{p(M_0|\mathcal{D})}}{\frac{p(M_1)}{p(M_0)}}$$

This is like a likelihood ratio, except we integrate out the parameters, which allows us to compare models of different complexity.

Bayes Factor $BF(1,0)$	**Interpretation**
$BF < \frac{1}{100}$	Decisive evidence for M_0
$BF < \frac{1}{10}$	Strong evidence for M_0
$\frac{1}{10} < BF < \frac{1}{3}$	Modest evidence for M_0
$\frac{1}{3} < BF < 1$	Weak evidence for M_0
$1 < BF < 3$	Weak evidence for M_1
$3 < BF < 10$	Modest evidence for M_1
$BF > 10$	Strong evidence for M_1
$BF > 100$	Decisive evidence for M_1

Jeffreys-Lindley paradox Problems can arise when we use improper priors (i.e. priors that do not integrate to 1) for model selection/ hypothesis testing, even though such priors may be acceptable for other purposes. In particular the Bayes Factor will always favor the simplest model since the probability of the observed data under a complex model with a very diffuse prior will be very small. Thus it is important to use proper priors when doing model selection.

2.3.4 Priors

The most controversial aspect of Bayesian statistics is its reliance on priors

Uninformative priors If we do not have strong evidence on what θ should be, it is common to use an uninformative priors, to "let the data speak for itself".
One might think that the most uninformative prior would be the uniform distribution: $Beta(1,1)$, but the posterior would then be: $\mathbb{E}(\theta|\mathcal{D}) = \frac{N_1+1}{N_1+N_0+2}$, whereas the MLE is $\frac{N_1}{N_1+N_0}$.
As by decreasing the magnitude of the pseudo counts, we can lessen the impact of the prior, we can argue that the most non-informative prior is:

$$\lim_{\epsilon \to 0} Beta(\epsilon, \epsilon) = Beta(0,0)$$

Called the *Haldane prior*, it is an improper prior.
In general it is advisable to perform a some kind of sensitivity analysis, in which one checks how much one's conclusions or prediction change in response to change in the modelling assumptions which includes the choice of the prior and the likelihood as well. If the conclusion are relatively insensitive to the modelling assumption, one can have more confidence in the results.

Jeffreys priors Harold Jeffreys designed a general purpose technique for creating non-informative priors. The key observation is that if $p(\phi)$ is non-informative then any re-parametrization of the prior, such as $\theta = h(\phi)$ for some function h should also be non-informative.

- Start with a variable change: $p_\theta(\theta) = p_\phi(\phi)\left|\frac{d\phi}{d\theta}\right|$
- Consider the following constraint: $p_\phi(\phi) \propto \sqrt{\mathcal{I}(\phi)}$, where $\mathcal{I}(\phi)$ is the Fisher information. $\mathcal{I}(\phi) \triangleq -\mathbb{E}\left(2 \times \frac{d\log(p(X|\phi))}{d\phi}\right)$. This a measure of the curvature of the expected negative log likelihood and hence a measure of stability of the MLE.

- Now $\frac{d\log(p(x|\theta))}{d\theta} = \frac{d\log(p(X|\phi))}{d\phi}\frac{d\phi}{d\theta}$
- $\mathcal{I}(\theta) = \mathcal{I}(\phi)\left(\frac{d\phi}{d\theta}\right)^2$
- $\sqrt{\mathcal{I}(\theta)} = \sqrt{\mathcal{I}(\phi)}\left|\frac{d\phi}{d\theta}\right|$
- Finally $p_\theta(\theta) = p_\phi(\phi)\left|\frac{d\phi}{d\theta}\right| \propto \sqrt{\mathcal{I}(\phi)}\left|\frac{d\phi}{d\theta}\right| = \sqrt{\mathcal{I}(\theta)}$

Robust priors To prevent an undue influence on the result, we build priors having heavy tails, which avoids forcing things to be too close to the prior mean.

Mixture of conjugate priors Conjugate priors simplify the computation of robust priors, but are often not robust, and not flexible enough to encode our prior knowledge. However it turns out that a mixture of conjugate priors is also conjugate, and seem to be a good compromise.

2.3.5 Hierarchical and Empirical Bayes

Hierarchical Bayes A key requirement for computing the posterior $p(\theta|\mathcal{D})$ is the specification of a prior $p(\theta|\eta)$ where η are the hyper-parameters. A Bayesian approach is to put a prior on our priors. This is an example of a **hierarchical Bayesian Model**.

Empirical Bayes In hierarchical Bayesian models, we need to compute the posterior on multiple levels of latent variables. For example, in a two-level model, we need to compute: $p(\eta, \theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta|\eta)p(\eta)$ We can approximate the posterior on the hyper-parameters with a point-estimate, $p(\eta|\mathcal{D} \approx \delta_{\hat{\eta}}(\eta))$ where $\hat{\eta} = \arg\max_\eta p(\eta|\mathcal{D})$. Since η is typically much smaller than θ in dimensionality, it is less prone to overfitting, so we can safely use a uniform prior on η. Then the estimate becomes:

$$\hat{\eta} = \arg\max_\eta p(\mathcal{D}|\eta) = \arg\max_\eta \int p(\mathcal{D}|\theta)p(\theta|\eta)d\theta$$

This overall approach is called **Empirical Bayes**
Empirical Bayes violates the principle that the prior should be chosen independently of the data. However, we can just view it as a computationally cheap approximation to inference in a hierarchical Bayesian model, just as we viewed MAP estimation as an approximation to inference in the one level model $\theta \to \mathcal{D}$. In fact, we can construct a hierarchy in which the more integrals one performs, the "more Bayesian" one becomes:

Method	**Definition**
Maximum likelihood	$\hat{\theta} = \arg\max_\theta p(\mathcal{D}|\theta)$
MAP estimation	$\hat{\theta} = \arg\max_\theta p(\mathcal{D}|\theta)p(\theta|\eta)$
ML-II (Empirical Bayes)	$\hat{\eta} = \arg\max_\eta \int p(\mathcal{D}|\theta)p(\theta|\eta)d\theta = \arg\max_\eta p(\mathcal{D}|\eta)$
MAP-II	$\hat{\eta} = \arg\max_\eta \int p(\mathcal{D}|\theta)p(\theta|\eta)p(\eta)d\theta = \arg\max_\eta p(\mathcal{D}|\eta)p(\eta)$
Full Bayes	$p(\theta, \eta|\mathcal{D}) \approx p(\mathcal{D}|\theta)p(\theta|\eta)p(\eta)$

2.3.6 Bayesian Decision Theory

We can formalize any given statistical decision problem as a game against nature (as opposed to a game against other strategic players, which is the topic of game theory). In this game, nature picks a state or parameter or label, $y \in \mathcal{Y}$, unknown to us, and then generates an observation, $\boldsymbol{x} \in \mathcal{X}$ which we get to see. We then have to make a decision, that is, we have to choose an action a from some **action space** $\mathcal{A}$. Finally we incur some **loss**, $L(y, a)$, which measures how compatible our action a is with nature's hidden state y.
Our goal is to devise a decision procedure or policy, $\delta : \mathcal{X} \to \mathcal{A}$ which specifies the optimal action for

each possible input which specifies the optimal action for each possible input, meaning the action that minimizes the expected loss:

$$\delta(\boldsymbol{x}) = \arg\min_{a \in \mathcal{A}} \mathbb{E}\left(L(y, a)\right)$$

In the Bayesian vision, the expected value of y given the data we have seen so far, whereas in the frequentist vision the expected value refers to x and y that we expect to see in the future.
In the Bayesian vision the optimal action having observed $\boldsymbol{x}$ is defined as the action a that minimizes the **posterior expected loss**:

$$\rho(a|\boldsymbol{x}) \triangleq \mathbb{E}_{p(y|x)}\left(L(y, a)\right) = \sum_y L(y, a) p(y|x)$$

Hence the Bayes estimator also called Bayes decision rule is given by:

$$\delta(\boldsymbol{x}) = \arg\max_{a \in \mathcal{A}} \rho(\boldsymbol{a}|\boldsymbol{x})$$

Bayes estimators for common loss functions

- **MAP** estimate minimizes 0-1 loss: $L(y, a) = \mathbb{I}_{y \neq a} \begin{cases} 0 \text{ if } a = y \\ 1 \text{ else} \end{cases}$
- **Reject option**, in classification problems where $p(y|\boldsymbol{x})$ is very uncertain we may prefer to choose a reject action, in which we refuse to classify the example as any of the specified classes. Let choosing $a = C + 1$ correspond to picking the reject action, and choosing $a \in \{1, ..., C\}$ correspond to picking one of the classes.
  $L(y = j, a = i) = \begin{cases} 0 & \text{if } i = j \text{ and } i, j \in \{1, ..., C\} \\ \lambda_r & \text{if } i = C + 1 \\ \lambda_s & \text{otherwise} \end{cases}$
  where λ_r is the cost of the reject action, and λ_s is the cost of a substitution error.
- **Squared Error (l_2)** for a continuous parameters. $L(y, a) = (y - a)^2$
- **Absolute Error (l_1)** more robust against outliers. $L(y, a) = |y - a|$. The optimal point is the median.
- **Supervised learning** considering a prediction function $\delta : \mathcal{X} \to \mathcal{Y}$ and some cost function $l(y, \delta(x))$. Then the loss incurred by taking action δ when the unknown state of nature is θ (the parameters of the data generating the mechanism). $L(\boldsymbol{\theta}, \delta) \triangleq \mathbb{E}_{(\boldsymbol{x}, y)\ p(\boldsymbol{x}, y|\boldsymbol{\theta})}\left(l(y, \delta(\boldsymbol{x}))\right) = \sum_{\boldsymbol{x}} \sum_y L\left(y, \delta(\boldsymbol{x}) p(\boldsymbol{x}, y|\boldsymbol{\theta})\right)$

Model evaluation metrics

- **False positive vs False negative trade-off** for binary decision problems three are 2 types of errors:
  1. false positive (false alarm) if $\hat{y} = 1 \wedge y = 0$
  2. false negative (missed detection) if $\hat{y} = 0 \wedge y = 1$

  We can consider the loss matrix:

Headers	$\boldsymbol{y = 1}$	$\boldsymbol{y = 0}$
$\boldsymbol{\hat{y} = 1}$	0	L_{FP}
$\boldsymbol{\hat{y} = 0}$	L_{FN}	0

  where L_{FN} is the cost of a false negative and L_{FP} the cost of a false positive.

- **ROC curves** From the below table

Headers		Truth		Count
Estimate	1	TP	FP	$\hat{N}_+ = TP + FP$
	0	FN	TN	$\hat{N}_- = FN + TN$
Count		$N_+ = TP + FN$	$N_- = FP + TN$	$N = N_+ + N_- = \hat{N}_+ + \hat{N}_-$

we can generate the *confusion matrix* is the below table

Headers	$\boldsymbol{y = 1}$	$\boldsymbol{y = 0}$
$\hat{\boldsymbol{y}} = \mathbf{1}$	$\frac{TP}{N}$ (sensitivity/recall)	$\frac{FP}{N}$ (error type I/ false alarm)
$\hat{\boldsymbol{y}} = \mathbf{0}$	$\frac{FN}{N}$ (error type II/ missed detection)	$\frac{TN}{N}$ (specificity)

- **Precision recall curves** When trying to detect a rare event the number of negatives is very large, hence comparing *sensitivity* and *the error of type I* is not very informative. We would then like to use a measure that only talks about positives.
  - **precision** $= \frac{TP}{\hat{N}_+}$
  - **recall** $= \frac{TP}{N_+}$

  A **precision recall curve** is a plot of *precision* vs *recall*.
- **F-scores** is the *harmonic mean of precision and recall*:
  $F_1 \triangleq \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$

2.4 Frequentist approach

2.4.1 Sampling distribution

Sampling Distributions of an estimator In frequentist statistic a parameter estimate $\hat{\theta}$ is computed by applying an estimator δ to some data $\mathcal{D}$, so $\hat{\theta} = \delta(\mathcal{D})$. The uncertainty in the parameter estimate can be measured by computing the *sampling distribution of the estimator*. Imagine sampling many different datasets $\mathcal{D}^{(s)}$ from some true model $p(\cdot|\theta^*)$ meaning $\mathcal{D}^{(s)} = \left\{x_i^{(s)} \hookrightarrow p(\cdot|\theta^*)\right\}_{1 \leq i \leq N}$ for $1 \leq s \leq S$ and θ^* is the true parameter. Now apply the estimator $\hat{\theta}(\cdot)$ to each $\mathcal{D}^{(s)}$ to get a set of estimates $\{\hat{\theta}(\mathcal{D}^{(s)})\}_{1 \leq s \leq S}$. As we late $S \to \infty$, the distribution induced on $\hat{\theta}(\cdot)$ is the **sampling distribution of the estimator**.

Bootstrap It is a simple *Monte Carlo* technique to approximate the sampling distribution. The idea is that if we knew the true parameters θ^*, we could generate S fake datasets of size N, from the true distribution. We could then compute our estimator from each sample, and use the empirical distribution of the resulting samples as our estimate of the sampling distribution.
Since θ is unknown, the idea of the **parametric bootstrap** is to generate the samples using $\hat{\theta}(\mathcal{D})$ instead. An alternative, called **non-parametric bootstrap** is to sample the x_i^s (with replacement) from the original data $\mathcal{D}$ and then compute the induced distribution as before.

2.4.2 Fequentist decision theory

In Frequentist decision theory there is a loss function and a likelihood, but there is no prior and hence no posterior or posterior expected loss. Thus there is no automatic way of deriving an optimal estimator, unlike the Bayesian case.
Instead, we are free to choose any estimator or decision procedure $\delta : \mathcal{X} \to \mathcal{A}$ we want.
Having chosen an estimator, we define its expected loss or risk as follows

$$R(\theta^*, \delta) \triangleq \mathbb{E}_{p(\tilde{\mathcal{D}}|\theta^*)}\left(L(\theta^*, \delta(\tilde{\mathcal{D}}))\right) = \int L\left(\theta^*, \delta(\tilde{\mathcal{D}})\right) p(\tilde{\mathcal{D}}) d\tilde{\mathcal{D}}$$

where $\tilde{\mathcal{D}}$ is data sampled from 'nature's distribution' which is represented by parameter θ^*. Whereas the Bayesian posterior expected loss:

$$p(a, \mathcal{D}, \pi) \triangleq \mathbb{E}_{p(\theta|\mathcal{D},\pi)}\left(L(\theta, a)\right) = \int_{\Theta} L(\theta, \boldsymbol{a}) p(\theta|\mathcal{D}, \pi) d\theta$$

We see that the Bayesian approach averages over θ, which is unknown, and conditions on $\mathcal{D}$ which is known. Unlike the frequentist approach averages over $\tilde{\mathcal{D}}$, thus ignoring the observed data, and conditions on θ^* which is unknown.

Bayes risk How to chose amongst the estimators? We need some way to convert $R(\theta^*, \delta)$ into single measure of quality, $R(\delta)$ which does not depend on knowing θ^*. One approach is to put a prior on θ^* and then to define **Bayes risk** of an estimator as follows:

$$R_B(\delta) \triangleq \mathbb{E}_{p(\theta^*)}\left(R(\theta^*, \delta)\right) = \int R(\theta^*, \delta) p(\theta^*) d\theta^*$$

A **Bayes estimator** or **Bayes decision rule** is one which minimizes the expected risk: $\delta_B \triangleq \arg\min_{\delta} R_B(\delta)$

Connection Bayesian and Frequentist approaches to decision theory.

- *Theorem 1* A Bayes estimator can be optained by minimizing the posterior expected loss for each $\boldsymbol{x}$
- *Theorem 2* Every admissible frequentist decision rule is a Bayes decision rule with respect to some possibly improper prior distribution.

Minimax risk Some frequentist statistic users avoid using Bayes risk since it requires the choice of a prior, although this is only in the evaluation of the estimator, not necessarily as part of its construction. An alternative approach is as follows:

1. Define the maximum risk of an estimator as:
   $R_{max}(\delta) \triangleq \max_{\theta^*} R(\boldsymbol{\theta}^*, \delta)$
2. A **minimax rule** is one which minimizes the maximum risk: $\delta_{MM} \triangleq \arg\min_{\delta} R_{\max}(\delta)$

Minimax estimators have a certain appeal, however computing them can be hard and furthermore they are very pessimistic. In most statistical situations, excluding games theoretic ones, assuming nature is an adversary is not a reasonable assumption.

Admissible estimators The basic problem with frequentis decision theory is that it relies on knowing the true distribution $p(\cdot|\theta^*)$ in order to evaluate the risk. However it might be the case that some estimators are worse than others regardless of the value of θ^*.
In particular if for $\theta \in \Theta, R(\theta, \delta_1) \leq R(\theta, \delta_2)$ bayesthen we say that δ_1 **dominates** δ_2.
An estimator is said to be **admissible** if it is not strictly dominated by any other estimator.
Admissibility is not enough

2.4.3 Desirable properties of estimators

Consistent estimators An estimator is said to be **consistent** if it eventually recovers the true parameters that generated the data as the sample size goes to infinity.

Unbiased estimator The **bias** of an estimator is defined as

$$bias\left(\hat{\theta}(\cdot)\right) = \mathbb{E}_{p(\mathcal{D}|\theta^*)}\left(\hat{\theta}(\mathcal{D}) - \theta^*\right)$$

The estimator is unbiased when the bias is equal to 0.

Minimum variance estimators A famous result called the **Cramerè-Rao lower bound** provides a lower bound on the variance of any unbiased estimator. More precisely: Let $(X_j)_{1\leq j\leq p} \hookrightarrow p(X|\theta_0)$ and $\hat{\theta}(\cdot)$ an unbiased estimator of θ^* Then, under various smoothness assumptions on $p(X|\theta_0)$ we have

$$\mathbb{V}(\hat{\theta}) \geq \frac{1}{nI(\theta^*)}$$

where $I(\theta^*)$ is the Fisher information matrix.

Bias-Variance Trade-off As $MSE = variance + bias^2$
It might be wise to use a biased estimator, so long as it reduces our variance, assuming our goal is to minimize squared error.

2.4.4 Empirical Risk Minimization

Frequentist issue Frequentist decision theory suffers from the fundamental problem that one cannot actually compute the risk function, since it relies on knowing the true data distribution. By contrast, the Bayesian posterior expected loss can always be computed since it conditions on the data rather that on θ^*.
However there is one setting which avoids this problem, it is when the task is to predict observable quantities, as opposed to estimating hidden variables or parameters.
Instead of looking at loss functions of the form $L(\boldsymbol{\theta}^*, \delta(\mathcal{D}))$ let us look at loss functions of the form $L(y, \delta(\boldsymbol{x}))$.
Then the risk becomes: $R(p_*, \delta) \triangleq \mathbb{E}_{(\boldsymbol{x},y)\hookrightarrow p_*}\left(L(y, \delta(\boldsymbol{x}))\right) = \sum_{\boldsymbol{x}}\sum_{\boldsymbol{y}} L(y, \delta(\boldsymbol{x}))p_*(\boldsymbol{x}, y)$ Where p_* represents "nature's distribution", indeed this distribution is unknown, but a simple approach is to use the empirical distribution, derived from some training data to approximate $p_*(x, y) \approx p_{emp}(x, y) \triangleq \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}(\boldsymbol{x})\delta_{y_i}(y)$ We define the empirical risk as follows:

$$R_{emp}(\mathcal{D}, \mathcal{D}) \triangleq R(p_{emp}, \delta) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \delta(x_i))$$

Regularized risk minimization

$$R'(\mathcal{D}, \delta) = R_{emp}(\mathcal{D}, \delta) + \lambda C(\delta)$$

where $C(\delta)$ measures the complexity of the prediction function $\delta(\boldsymbol{x})$ and λ controls the strength of the complexity penalty. This approach is known as **regularized risk minimization**.

2.4.5 Components

Introduction Avoid treating parameters as randome variables. The notion of variation across repeated trials forms the basis for modelling uncertainity.

Hypothesis Testing A *frequentist* statistics, probabilities represent the frequencies at which particular events happen.

p-value It is the heart of frequentist hypothesis testing, it tells us the probability of getting a particular test statistic t as big as the one we have or bigger under the null hypothesis (that there is actually no effect).
By convention we usually conclude an effect is *statistically significant* if the *p-value* is less than a threshold α.

Confidence intervals When we fit a model to our data we look for the *maximum of likelihood* parameters, meaning the parameters that are most consistent with our data. For each parameter we will able to construct 95% interval namely 95 of the 100 intervals generated will contain the true value of the parameter.
If $H_0 : \beta = 0$ is true, the probability of getting a 95% confidence interval that does not include 0 is less than 0.05. In other words, if the 95% confidence does not include 0, $p < 0.05$.

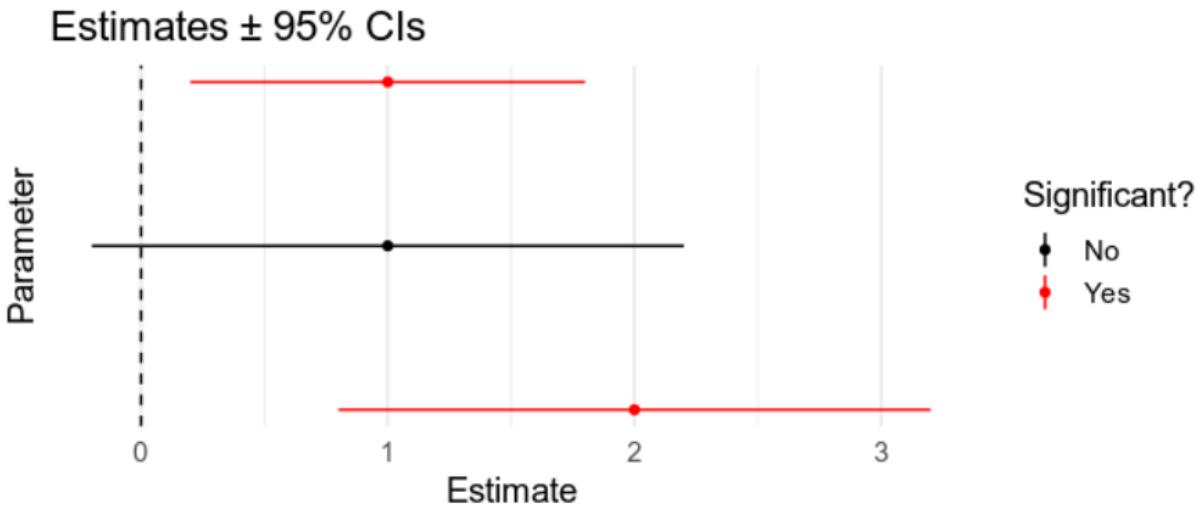

Figure 2.3: Confidence interval

Multiple comparisons The more tests we run the more likely it is to we'll find at least one that is significant even though the null hypothesis is true. We can then apply a Bonferroni correction.
Let's say we are running k tests, we can either adjust:

- the threshold $\alpha_{adj} = \frac{\alpha}{k}$ OR
- the *p-value* $p_{ajd} = k \times p$

Chapter 3

Conventional Statistical Learning

Chapter 4

Deep Learning

Chapter 5

Use-cases

Bibliography

[1] Omar Elgabry. *The Ultimate Guide to Data Cleaning*. 2019. URL: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4.

[2] Wikipedia contributors. *Moments (Mathematics)*. [Online; accessed 21-August-2023]. 2023. URL: https://en.wikipedia.org/wiki/Moment_(mathematics).

[3] Wikipedia contributors. *Probability*. [Online; accessed 20-August-2023]. 2023. URL: https://en.wikipedia.org/wiki/Probability.