SUMMARY ON STATISTICS

1. Useful probability distributions

1.1) Binary variables

$$y = \underline{\{0,1\}}, \ \{-1,1\}, \ \{A, B\}, \ldots$$

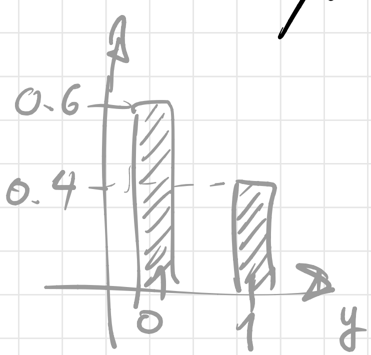
(the way to treat a binary classification problem)

$$y \sim \text{Bern}(y|\mu)$$

"pipe" notation ⇏ conditioned to ...

$$= \mu^{y}(1-\mu)^{1-y}$$

the likelihood of depending on parameter μ



$$\mu = p(y=1)$$

$$p(y=0) = 1-\mu$$

Imagine we have a dataset $D = \{\vec{x}_i, y_i\}_{i=1}^{N}$

e.g. y_i : type of galaxy
(spiral or elliptical)
$\vec{x}_i$: position in the sky [[given]]

for i.i.d. data, the total likelihood is :

$$p(\vec{y} \mid X, \vec{\theta}) = \prod_{i=1}^{N} \mathrm{Bern}\left(y_i \mid \mu(\vec{x}_i; \vec{\theta})\right)$$

$$= \prod_i \mu(\vec{x}_i; \vec{\theta})^{y_i} \left(1 - \mu(\vec{x}_i; \vec{\theta})\right)^{1-y_i}$$

$\mu(\vec{x}; \vec{\theta}) \Rightarrow$ This is given by our model
(ML model, physics model, ...)

Given a model for μ, we need to fit it to our data, i.e. to find the optimum parameters, $\vec{\theta}_{opt}$.

e.g. $\vec{\theta}_{opt} = \underset{\vec{\theta}}{\mathrm{argmax}}\; p(\vec{y} \mid X, \vec{\theta})$

↳ Maximum Likelihood Estimator

1.2) Categorical variables

{ e.g. types of jets at LHC, types of events at particle detectors }

$$y = \{1, 2, \dots, k\}, \quad \{A, B, C, \dots\}$$

parametrise it in a computer-convenient way: "1-to-K scheme" or "one-hot-encoding":

Suppose $k=3$ classes

$$\text{if} \quad y = 1 \rightarrow \vec{y} = \{1, 0, 0\}$$
$$y = 2 \rightarrow \vec{y} = \{0, 1, 0\}$$
$$y = 3 \rightarrow \vec{y} = \{0, 0, 1\}$$

i.e. $y = k \rightarrow \vec{y} = \{0, 0, \dots, 1, 0, 0, \dots\}$

k^{th} position

These variables are distributed according to: the categorical distribution (or generalized Bernoulli):

$$p(\vec{y} \mid \vec{\mu}) = \prod_{k=1}^{K} \mu_k^{y_k}$$

$\mu_k = P(y_k = 1)$

$\{ \vec{\mu} = \{\mu_k\} \}$

indeed for $K=2$ (binary case)

$$p(\vec{y}\,|\,\vec{\mu}) = \mu_1^{y_1}\mu_2^{y_2} = \mu_1^{y_1}(1-\mu_1)^{1-y_1}$$

$$\mu_1 = p(y_1=1) = p(y_2=0)$$

$$= \mathrm{Bern}(y_1\,|\,\mu_1)$$

So what is the likelihood of a dataset of categorical variables?

one-hot-encoded

$$D = \{\vec{x}_i, \vec{y}_i\}_{i=1}^{N}$$

$$X = \{\vec{x}_i\},\ Y = \{\vec{y}_i\}$$

$$P(Y\,|\,X,\vec{\theta}) = \prod_{i=1}^{N}\prod_{k=1}^{K}\mu(\vec{x}_i,\vec{\theta})^{y_{ik}}$$

$$\underbrace{-\ln p(Y\,|\,X,\vec{\theta}) = -\sum_i\sum_k y_{ik}\ln\mu(\vec{x}_i,\vec{\theta})}$$

↳ "Cross entropy" cost function

↳ to minimise to obtain $\vec{\theta}_{opt}$

1.3) Integer-valued variables

$y = \{0, 1, 2, \dots\} \quad y \in \mathbb{Z}$

any counting experiment, photons from the sky, etc

(not really a categorical variable)

- Under certain assumptions (independence, non-overlapping, small rate)

$$y \sim \mathrm{Pois}(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

Poisson distribution with mean λ

So assume having data:

$$D = \{\vec{x}_i, y_i\}_{i=1}^{N}$$

$$p(\vec{y}|X;\vec{\theta}) = \prod_{i=1}^{N} \frac{1}{y_i!} \lambda(\vec{x}_i;\vec{\theta})^{y_i} \cdot \exp\{-\lambda(\vec{x}_i;\vec{\theta})\}$$

1.4) Continuous variables

$y \in \mathbb{R}$ { Flux of particles, energy deposition, time of flight, ... }

The simplest, most common, is to assume a Gaussian distribution

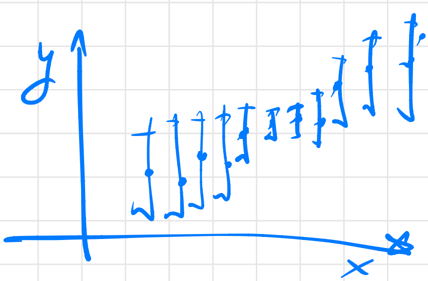
$$y \sim \mathcal{N}(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}$$

or its multivariate version

$$\vec{y} \sim \mathcal{N}(\vec{y} \mid \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}(\det\Sigma)^{1/2}} \cdot \exp\left\{-(\vec{y}-\vec{\mu})^T \cdot \Sigma^{-1} \cdot (\vec{y}-\vec{\mu})\right\}$$

Again the aim is to model μ.

e.g.



$\Rightarrow$ Fit a straight line to these data

$$D = \{x_i, y_i, \sigma_i\} \qquad y_i \sim \mathcal{N}(y_i \mid \mu(x_i; \vec{\theta}), \sigma_i^2)$$

$$\mu(x, \vec{\theta}) = \theta_1 x + \theta_2$$

The data likelihood :

$$p(\vec{y} \mid \vec{x}, \vec{\theta}) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid \mu(x_i, \vec{\theta}), \sigma_i^2)$$

$$\log p(\vec{y} \mid \vec{x}, \vec{\theta}) = \sum_i - \frac{(y_i - \theta_1 x_i - \theta_2)^2}{2\sigma_i^2} + \text{const}$$

$$\vec{\theta}_{opt} = \underset{\vec{\theta}}{\operatorname{argmin}} \sum_i \frac{(y_i - \theta_1 x_i - \theta_2)^2}{\sigma_i^2}$$ } Least squares procedure!

The results for θ_1 and θ_2 coming from this minimisation coincide with the one in textbooks

All this belongs to the "frequentist" approach, Outcome is a priori $\vec{\theta}_{opt}$, i.e. a point-like prediction without uncertainty estimation.

There are several methods to estimate the uncertainty on $\vec{\theta}_{opt}$ in the frequentist framework

Bootstrap, Analytical method, Fisher information... or just Neyman–Pearson intervals

On the other hand, uncertainties can be estimated "from first principles" using the Bayesian approach:

$$D = \{\vec{x}_i, y_i\}_{i=1}^{N}$$

$p(\vec{y}\,|\,X, \vec{\theta})$ is the same likelihood as before

$$p(\vec{\theta}\,|\,D) = \frac{p(D\,|\,\vec{\theta})\, p(\vec{\theta})}{p(D)}$$

prior distribution of the parameters of the model "m"

"Evidence" or "marginal likelihood"

$$p(D) = p(D\,|\,m) = \int d\vec{\theta}\; p(D\,|\,\vec{\theta})\, p(\vec{\theta})$$

Since $p(D)$ doesn't depend on $\vec{\theta}$, it is not needed for computing $\vec{\theta}_{opt}$ given a dataset (however it is needed for "model comparison", as we will see later)

- So $\vec{\theta}_{opt}$ is obtained in this framework by maximising the (un-normalised) posterior

$$\vec{\theta}_{opt} = \vec{\theta}_{MAP} = \underset{\vec{\theta}}{\operatorname{argmax}}\ p(D|\vec{\theta})\,p(\vec{\theta})$$

So if we have a way to estimate $p(\vec{\theta}|D)$, then it naturally gives a way to estimate the uncertainties

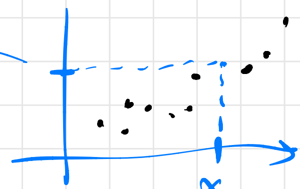
$$1-\beta = \int_{\vec{\theta}_{min}}^{\vec{\theta}_{max}} d\vec{\theta}\, p(\vec{\theta}|D) = \underbrace{\overset{\text{e.g.}}{0.95}}_{\text{e.g. 95\% credible interval for } \beta=0.05}$$

$$\vec{\theta} = [\vec{\theta}_{min}, \vec{\theta}_{max}]$$

- In many situations we need to go a step beyond $\vec{\theta}_{opt}$ and actually make a prediction for a new input $\vec{x}_{new}$

  - Frequentist approach:
    Simply substitute $\vec{\theta}_{opt}$ inside the expected value of "y":

$$\mu(\vec{x}_{new}; \vec{\theta}_{opt})$$



  - Bayesian approach:
    Distribution of $\vec{\theta}$ $p(\vec{\theta}|D)$ translates into a distribution of the prediction, a.k.a. the "predictive distribution"

$$p(y_{new}|\vec{x}_{new}, D) = \int d\vec{\theta}\, p(\vec{\theta}|D) \underbrace{p(y_{new}|\vec{x}_{new}, \vec{\theta})}_{\text{likelihood of } y_{new}}$$

In very few situations $p(\vec{\theta} | D)$ is analytical (or even tractable!).

The prototypical example is in the following situation:

- Prior: $p(\vec{\theta}) = \mathcal{N}(\vec{\theta} \,|\, \vec{\mu}_0, \Sigma_0)$
- Likelihood: $p(\vec{y} | \vec{\theta}) = \mathcal{N}(\vec{y} \,|\, \Phi(x) \cdot \vec{\theta} + \vec{b}, C)$

i.e. Both prior & likelihood are Gaussian, but the mean of the likelihood <u>has to be</u> linear in the parameters!

(e.g. a polynomial:

$$f(x; \vec{\theta}) = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots,$$

but also a linear combination of non-linear functions of x:

$$f(x; \vec{\theta}) = \sum_{k=0}^{M} \theta_k \cdot \phi_k(x) \equiv \vec{\theta}^T \cdot \vec{\phi}(x)$$

$$\vec{\phi}(x) = \{1, \phi_1(x), \phi_2(x), \ldots, \phi_M\}$$

$\{\phi_k(x)\}_{k=1}^{M}$ is a given catalog of non-linear functions (whose params. are fixed)

So if we have N input values $\{x_i\}$ we can build a matrix

$$\Phi(X) = \begin{pmatrix} 1 & \phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_M(x_1) \\ 1 & \phi_2(x_1) & \phi_2(x_2) & \cdots & \phi_M(x_2) \\ \vdots & \vdots & & & \\ 1 & \phi_1(x_N) & \phi_2(x_N) & \cdots & \phi_M(x_N) \end{pmatrix}$$

Note that in spite of having a simple structure, this is model is actually a "universal approximator", the same way a neural network is! With the advantage that the solution is analytical!