Summary on Statistics

1) Useful probability distrib.

1.1 Binary variables

$$x = \{0, 1\}$$

Let $p(x=1) = \mu$

↳ $\text{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$

↳ Basis of binary classification

Dataset $D = \{x_i\}_{i=1}^N$ (i.i.d)

↓

Likelihood $p(D|\mu) = \prod_{i=1}^N \text{Bern}(x_i|\mu)$

↳ object of interest

1.2. Categorical variables

$$X = \{1, 2, \dots, K\}$$

↳ parametrise it in the 1-to-K scheme
"one-hot encoding"

A point $x_i \rightarrow \vec{x}_i = \{x_i^1, x_i^2, \dots, x_i^K\}$

↳ binary vector

e.g. $K = 5$

$x_i = 3 \rightarrow \vec{x}_i = (0, 0, 1, 0, 0)$

Distrib. is the Categorical distrib.
(or generalized Bernoulli)

$$g_{Bern}(\vec{x} \mid \vec{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$\vec{\mu} = \{\mu_k\}_{k=1}^{K}$

e.g. for $K = 2$ classes

$g_{Bern} = \mu_1^0 \mu_2^1$ if $\vec{x} = (0, 1)$

$= \mu^0 (1-\mu)^1 \Rightarrow$ Bernoulli

$\mu_1 + \mu_2 = 1$

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

PDF:

$$p(D|\vec{\mu}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{ik}}$$

$$-\ln p(D|\vec{\mu}) = -\sum_{i=1}^{N} \sum_{k=1}^{K} x_{ik} \ln \mu_k$$

↳ "cross-entropy" loss function

1.3) Continuous variables

Gaussian for a multivariate point $\vec{x}$ ↶ dim = D

$$N(\vec{x}|\vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}}$$

$$\times \exp\left\{ -\frac{1}{2} (\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu}) \right\}$$

covariance

Σ = DxD matrix

$$\Sigma = \begin{pmatrix} \mathbb{E}[(x_1-\mu_1)(x_1-\mu_1)] & \cdots \\ \mathbb{E}[(x_2-\mu_2)(x_1-\mu_1)] & \cdots \\ \vdots & \\ \mathbb{E}[(x_D-\mu_D)(x_1-\mu_1)] & \cdots \end{pmatrix}$$

2.) Frequentist approach to inference

Likelihood of data D

$$p(D|\mu) = \prod_{i=1}^{N} p(x_i|\mu)$$

The aim is to model μ

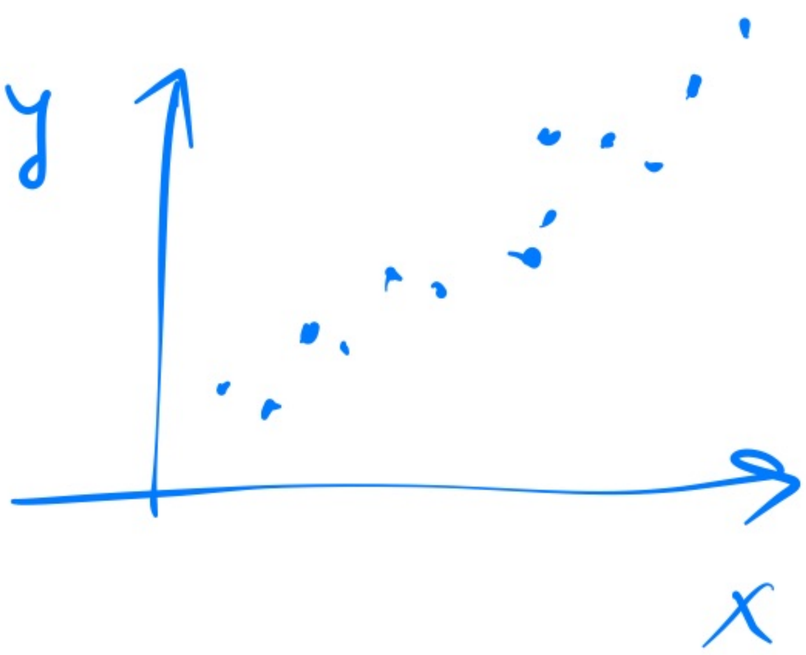
$$\mu = h(\vec{w})$$

↳ predefined function

$$\vec{w}_{MLE} = \text{argmax}\ p(D|\mu) = \mathcal{L}(\vec{w})$$

For example $D = \{x_i, y_i\}_{i=1}^{N}$

↳ output



$$y(x) \sim \mathcal{N}(y(x)|\mu(x,\vec{w}), \sigma^2)$$

let's model

$$\mu(\vec{x},\vec{w}) = ax + b$$

$$\vec{w} = \{a, b\} \qquad f(x,\vec{w})$$

Likelihood

$$\mathcal{L}(a,b) = \prod_{i=1}^{N} \mathcal{N}\left(y_i \mid f(x_i, a, b), \sigma^2\right)$$

$$-2 \ln \mathcal{L} = -2 \sum_i \left\{ \ln \frac{1}{\sqrt{2\pi}} + \frac{1}{2} \ln \frac{1}{\sigma^2} - \frac{(y_i - a x_i - b)^2}{2\sigma^2} \right\}$$

maximise $\mathcal{L}$

equivalent to minimise

$$\Rightarrow \quad \epsilon = \sum_i (y_i - a x_i - b)^2$$

Obtain ↳ least squares fitting procedure

↳

$$a^* = \frac{N \sum_i y_i x_i - \left(\sum_i x_i\right)\left(\sum_i y_i\right)}{N \sum_i x_i^2 - \left(\sum_i x_i\right)^2}$$

$$b^* = \frac{1}{N}\left(\sum_i y_i - a^* \sum_i x_i\right)$$

Once we have $\vec{w}^*_{MLE}$

⇒ predictor for a new point

$$p(y_{new} \mid \vec{x}_{new}, \vec{\omega}_{MLE})$$

$$= \mathcal{N}(y_{new} \mid a^* x_{new} + b^*, \sigma^2)$$

point-like prediction without associated uncertainties

⇓

a way out ⇒ bootstrapping

⇓

Bayesian approach

The final aim typically is the estimation of predictive distribution

$$p(y_{new} \mid x_{new}, \vec{D})$$

$\vec{D}$ → $\{\vec{x}, \vec{y}\}$ training

$$= \int d\vec{\omega}\; p(y_{new} \mid \vec{x}_{new}, \vec{\omega})\; p(\vec{\omega} \mid D)$$

likelihood of y_{new}

posterior

↓ one can compute the 2nd moment in order to estimate the uncertainties of the prediction

- Assume some data about observable $\vec{y}$

  $\Rightarrow$ take it as following a Gaussian distrib.

$$\Rightarrow p(\vec{y}|\vec{w}) = \mathcal{N}(\vec{y}|X\cdot\vec{w}+\vec{b}, L^{-1})$$

↳ inverse of the cov. matrix

"precision matrix"

- Assume that $\vec{w}$ follow a Gaussian as well } prior on $\vec{w}$

$$\Rightarrow p(\vec{w}) = \mathcal{N}(\vec{w}|\vec{\mu}, \Lambda^{-1})$$

question: $p(\vec{w}|\vec{y})$? (posterior)

$p(\vec{y})$? ("evidence")

⇓ assignment

Answer: $p(\vec{y}) = \mathcal{N}(\vec{y}|X\vec{\mu}+\vec{b}, L^{-1} + X\Lambda^{-1}X^T)$

$p(\vec{w}|\vec{y}) = \mathcal{N}(\vec{w}|\Sigma\{X^TL(\vec{y}-\vec{b}) + \Lambda\vec{\mu}, \Sigma)$

cov. of

$\Sigma = (\Lambda + X^TLX)^{-1}$