Probabilistic modeling

Probabilities

Motivation: many machine learning problems can be conveniently framed into a probabilistic framework

Quick refresher on probabilites

- A **random variable** X represents the uncertainty of a certain event
- X can be either **discrete** (finite number of possible values) or **continuous** (real values)
- X is described by:
  - (discrete) its probability mass function $P(X = x)$
  - (continuous) its probability density function p such that $P(a \leqslant X \leqslant b) = \int_a^b p(x)\mathrm{d}x$
- A **multivariate** random variable $\boldsymbol{X}$ in dimension D is a vector of random variables, taking values $\boldsymbol{x} = [x_1, \ldots, x_D]^\mathsf{T}$

Sum rule, product rule, Bayes theorem

- **Joint** probability[1] of two random variables: $p(x, y)$
- **Sum rule**: the marginal probability is given by

$$p(x) = \begin{cases} \sum_y p(x, y) & \text{if } y \text{ is discrete} \\ \int p(x, y) \mathrm{d}y & \text{if } y \text{ is continuous} \end{cases}$$

- **Product rule**: relates joint and conditional probabilities

$$\begin{aligned} p(x, y) &= p(y \mid x)\, p(x) \\ &\quad (= p(x \mid y)\, p(y)) \end{aligned}$$

- From the product rule, we can derive **Bayes theorem**:

$$\underbrace{p(y \mid x)}_{\text{posterior}} = \frac{\overbrace{p(x \mid y)}^{\text{likelihood}}\ \overbrace{p(y)}^{\text{prior}}}{\underbrace{p(x)}_{\text{evidence}}}$$

[1] for continuous variables, we actually refer to the probability density function

Expected value and variance

For a random variable X,

- The **expected value** of a function f over x is defined as:

$$\mathbb{E}_{\mathrm{x}}[f] = \sum_{x} p(x) f(x) \qquad \text{(in the discrete case)}$$

$$\mathbb{E}_{\mathrm{x}}[f] = \int p(x) f(x)\, \mathrm{d}x \qquad \text{(in the continuous case)}$$

- The **variance** of f is then defined as:

$$\mathrm{Var}_{\mathrm{x}}[f] = \mathbb{E}_{\mathrm{x}}\left[\, (f(x) - \mathbb{E}_{\mathrm{x}}[f])^2 \,\right]$$

For a multivariate random variable $\boldsymbol{X}$ in dimension D,

- The expected value is a vector $\mathbb{E}_{\boldsymbol{X}}[\boldsymbol{x}] = [\mathbb{E}[x_1], \ldots, \mathbb{E}[x_D]]^\mathsf{T}$

- The covariance matrix $\mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{x}]$ is the symetric matrix of co-variances of its components

Bernoulli distribution

A discrete random variable X with states $x \in \{0, 1\}$ such that:

$$p(x) = \mu^x (1 - \mu)^{1-x}$$

where $\mu \in [0, 1]$ is the parameter of the distribution.

We have:

$$\mathbb{E}[x] = \mu$$
$$\mathrm{Var}[x] = \mu(1 - \mu)$$

Example: flip of a coin.

Multivariate case: **categorical** or **multinoulli** distribution over k states, parameterized by vector $\boldsymbol{p} \in [0, 1]^k$ where $\sum_i p_i = 1$.

Normal distribution (gaussian)

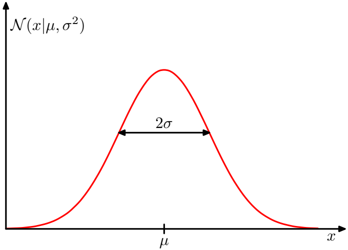
A continuous random variable x following a normal distribution has the following probability density function:

$$p(x) = \mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Parameterized by its mean μ and variance σ^2:

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2)\, x \,\mathrm{d}x = \mu$$

$$\mathrm{Var}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2)\, (x-\mu)^2 \,\mathrm{d}x = \sigma^2$$



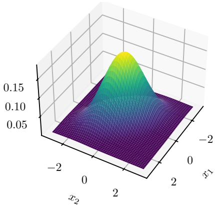
Multivariate normal distribution

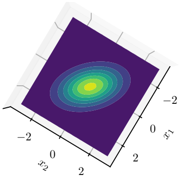
For continuous random vector $\mathbf{x}$ in dimension d, the probability density function is given by:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right)$$

Parameterized by its mean vector $\boldsymbol{\mu}$ and its covariance matrix $\boldsymbol{\Sigma}$:

$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu} \qquad \mathrm{Var}[\boldsymbol{x}] = \boldsymbol{\Sigma}$$



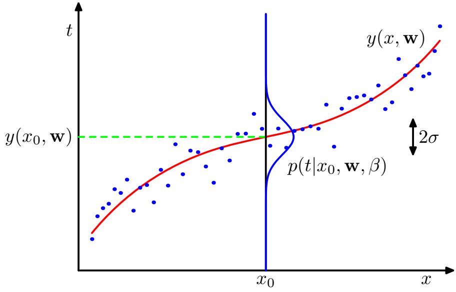


Linear regression as a probabilistic model

Recall the linear regression problem: $y = f(x; \boldsymbol{w})$

Hypothesis: assume target y is generated from a normal distribution with a mean equal to $f(x; \boldsymbol{w})$

$$p(y \mid x, \boldsymbol{w}, \sigma^2) = \mathcal{N}(y \mid f(x; \boldsymbol{w}), \sigma^2)$$



Also assume y_i are drawn independently from each other (conditional independence).

Maximum likelihood estimation

With target vector $\boldsymbol{y}$ and data matrix $\boldsymbol{X}$, we can write the conditional likelihood of the training data:

$$p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid (f(\boldsymbol{x}_i, \boldsymbol{w}), \sigma^2)$$

Maximum likelihood (ML) estimation: find $\boldsymbol{w}$ by maximizing the conditional (log-)likelihood of the training data

$$\boldsymbol{w}_{\mathrm{ML}} = \arg\max_{\boldsymbol{w}} \; \log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}, \sigma^2)$$

with:

$$\log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}, \sigma^2) = -\frac{N}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2$$

$\rightarrow$ **Equivalent to minimizing the mean squared error**

Maximum a posteriori estimation

Hypothesis 2: parameters $\boldsymbol{w}$ follow a prior normal distribution

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{0}, b^2\boldsymbol{I})$$

Maximum a posteriori (MAP) estimation:

$$\boldsymbol{w}_{\text{MAP}} = \arg\max_{\boldsymbol{w}} \; \log p(\boldsymbol{w} \,|\, \boldsymbol{X}, \boldsymbol{y}) = \arg\max_{\boldsymbol{w}} \; \log \frac{p(\boldsymbol{y} \,|\, \boldsymbol{X}, \boldsymbol{w}) p(\boldsymbol{w})}{p(\boldsymbol{y}, \boldsymbol{X})}$$

We have:

$$\begin{aligned} \log p(\boldsymbol{w} \,|\, \boldsymbol{X}, \boldsymbol{y}) &= \log p(\boldsymbol{y} \,|\, \boldsymbol{X}, \boldsymbol{w}) + \log p(\boldsymbol{w}) + \text{const} \\ &= -\frac{1}{\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 - \frac{1}{2b^2} \|\boldsymbol{w}\|^2 + \text{const} \end{aligned}$$

$\rightarrow$ **Equivalent to minimizing the regularized mean squared error**