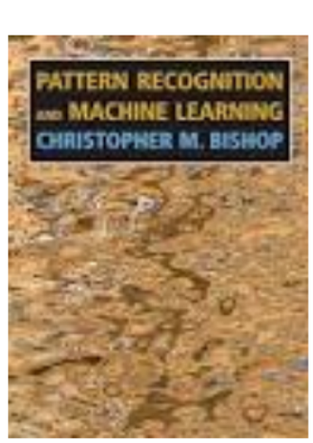
Uncertainty Quantification

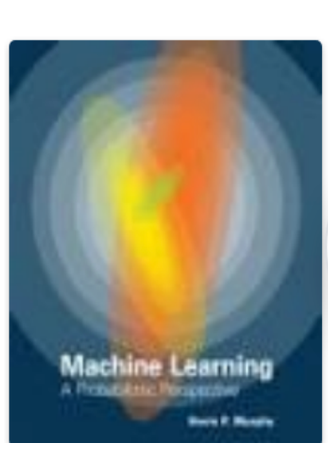
Sep 2022

Uncertainty Quantification

Recommended for reading:

- Bishop, Pattern Recognition and Machine Learning
- Murphy, Machine Learning A probabilistic approach
- These notes (from various sources)





Prerequisites:

Probability

Numerical methods

Knowledge of deep learning

Participation 20%

Final project 80%

Uncertainty Quantification

- Introduction
  - Uncertainty and predictive modeling
  - RV and distribution
  - MLE and statistical decision
  - Bayesian perspective

- Simulation
  - RV generation and CLT
  - Monte Carlo estimation
  - Laplace approximation and adjoint method

Uncertainty Quantification

- Regression
  - Linear and logistic regression
  - Generalized linear model
  - Bayesian linear regression
  - Gaussian prior and Gaussian regression

- Bayesian modeling
  - Bayesian learning
  - Prediction with posterior distribution
  - Bayesian neural networks

Uncertainty Quantification

- Probabilistic graphical models
  - Bayesian network
  - Directed and undirected graphical models
  - Hidden Markov model and EM
  - Exponential families

- Further topics
  - MCMC and sequencial Monte Carlo
  - Variational inference

Uncertainty Quantification

- Probabilistic graphical models
  - Bayesian network
  - Directed and undirected graphical models
  - Hidden Markov model and EM
  - Exponential families

- Further topics
  - MCMC and sequencial Monte Carlo
  - Variational inference

Probability

定义 4.8 设 $\mathscr{F}$ 是由 Ω 的一些子集组成的集合(这种由集合组成的集合一般叫做集合系),$P=P(\cdot)$是 $\mathscr{F}$ 上有定义的实值函数.若定义域 $\mathscr{F}$ 和函数 P 满足下列条件:

(1) $\Omega\in\mathscr{F}$; (4.1)

(2) 若 $A\in\mathscr{F}$,则 $A^c=\Omega-A\in\mathscr{F}$; (4.2)

(3) 若 $A_n\in\mathscr{F}$(一切 $n\geqslant 1$),则 $\bigcup_{n=1}^{\infty}A_n\in\mathscr{F}$; (4.3)

(4) $P(A)\geqslant 0$ (一切 $A\in\mathscr{F}$); (4.4)

(5) $P(\Omega)=1$; (4.5)

(6) 若 $A_n\in\mathscr{F}$(一切 $n\geqslant 1$),且两两不相交,就有

$$P\Big(\bigcup_{n=1}^{\infty}A_n\Big)=\sum_{n=1}^{\infty}P(A_n), \tag{4.6}$$

则称 P 是 $\mathscr{F}$ 上的**概率测度**(简称**概率**),$P(A)$为 A 的**概率**(也称 A 发生的概率).(见注)

Conditional probability

Two children, the older child is a girl.
What is the probability that both children are girls?

Two children, at least one of them is a boy.
What is the probability that both children are boys?

Conditional probability

Two children, the older child is a girl.
What is the probability that both children are girls?

Two children, at least one of them is a boy.
What is the probability that both children are boys?

$$\frac{\#\{girl\text{-}girl\}}{\#\{boy\text{-}girl, girl\text{-}girl\}} = \frac{1}{2}$$

Conditional probability

Two children, the older child is a girl.
What is the probability that both children are girls?

Two children, at least one of them is a boy.
What is the probability that both children are boys?

$$\frac{\#\{girl\text{-}girl\}}{\#\{boy\text{-}girl, girl\text{-}girl\}} = \frac{1}{2}$$

$$\frac{\#\{boy\text{-}boy\}}{\#\{boy\text{-}girl, girl\text{-}boy, boy\text{-}boy\}} = \frac{1}{3}$$

Conditional probability

定义 5.1 设 A 和 B 都是条件 S 下的事件，则称条件 S 已经实现且 B 已发生的情形下 A 发生的概率为 B 发生的条件下 A 的**条件概率**，并记为 $P(A|B)$.

一般情形下，怎样计算条件概率呢？有下列计算公式：

$$P(A|B) = \frac{P(AB)}{P(B)} \quad (\text{当 } P(B) \neq 0).$$

$$P(AB) = P(B) \cdot P(A|B).$$

Product formula

定理 5.1(一般乘法公式) 设 $A_1, A_2, \cdots, A_n$ 是 n 个事件($n \geqslant 2$),满足 $P(A_1A_2\cdots A_{n-1}) \neq 0$,则

$$P(A_1\cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots \cdot P(A_n|A_1\cdots A_{n-1}). \tag{5.3}$$

定义 5.2 设 A 和 B 都是条件 S 下的随机事件,若满足

$$P(AB) = P(A)P(B),$$

则称 A 与 B **相互独立**.("相互独立"简称"独立".)

Law of total probability

定理 6.1(全概公式)[①] 如果事件组 $B_1, B_2, \cdots, B_n (n \geqslant 2)$ 满足下列条件：

(1) $B_1, B_2, \cdots, B_n$ 两两不相容且 $P(B_i) > 0 (i = 1, 2, \cdots, n)$；

(2) $B_1 \cup B_2 \cup \cdots \cup B_n$ 是必然事件，

则对任何事件 A 皆有

$$P(A) = \sum_{i=1}^{n} P(B_i) P(A \mid B_i). \tag{6.2}$$

Bayesian formula

定理 6.2(逆概公式) 如果事件组 $B_1,\cdots,B_n(n\geqslant 2)$ 满足下列条件：

(1) $B_1,\cdots,B_n$ 两两不相容且 $P(B_i)>0(i=1,2,\cdots,n)$；

(2) $\bigcup_{i=1}^{n} B_i$ 是必然事件，

则对任一事件 A，只要 $P(A)>0$，就有

$$P(B_k|A)=\frac{P(B_k)P(A|B_k)}{\sum_{i=1}^{n}P(B_i)P(A|B_i)}. \tag{6.6}$$

Bayesian formula

H represents our **hypothesis** we wish to examine.

E represents **evidence** for that hypothesis (presumably evidence we obtain through experiment and/or observation).

P(H|E) is called the **posterior**. This value gives us the probability of our hypothesis given our obtained evidence. In a sense, it shows the degree that your hypothesis is confirmed from your evidence. This is the unknown end value that we would wish to obtain.

https://www.psychologyinaction.org/psychology-in-action-1/2021/10/19/whats-all-the-buzz-about-bayes

Bayesian formula

P(H) is called the **prior**. This is how probable our hypothesis was before any evidence is taken into account. The prior is infamous for being hard to define properly. How could you know any prior information about a hypothesis you haven't even tested? Without going into too much detail (because one could go on indefinitely about Bayesian priors), just know that this value has to be assumed. There are methods that can pick out better prior probabilities amongst worse ones, such as examining the results of previously conducted experiments. The prior that is developed from past experiments, however, will still be arbitrary to a degree.

P(E) is called the marginal likelihood/probability or often just the **marginal**. This value represents the probability of our evidence given all possible hypotheses. Most of the time, we do not actually need a direct value of the marginal, as it can often be reduced out of the equation.

https://www.psychologyinaction.org/psychology-in-action-1/2021/10/19/whats-all-the-buzz-about-bayes

随机硬币

两枚硬币，一枚均匀，另一枚正面概率为3/4。现随机取一枚硬币，连续投掷3次都是正面！取到的是均匀硬币的概率多大?

随机硬币

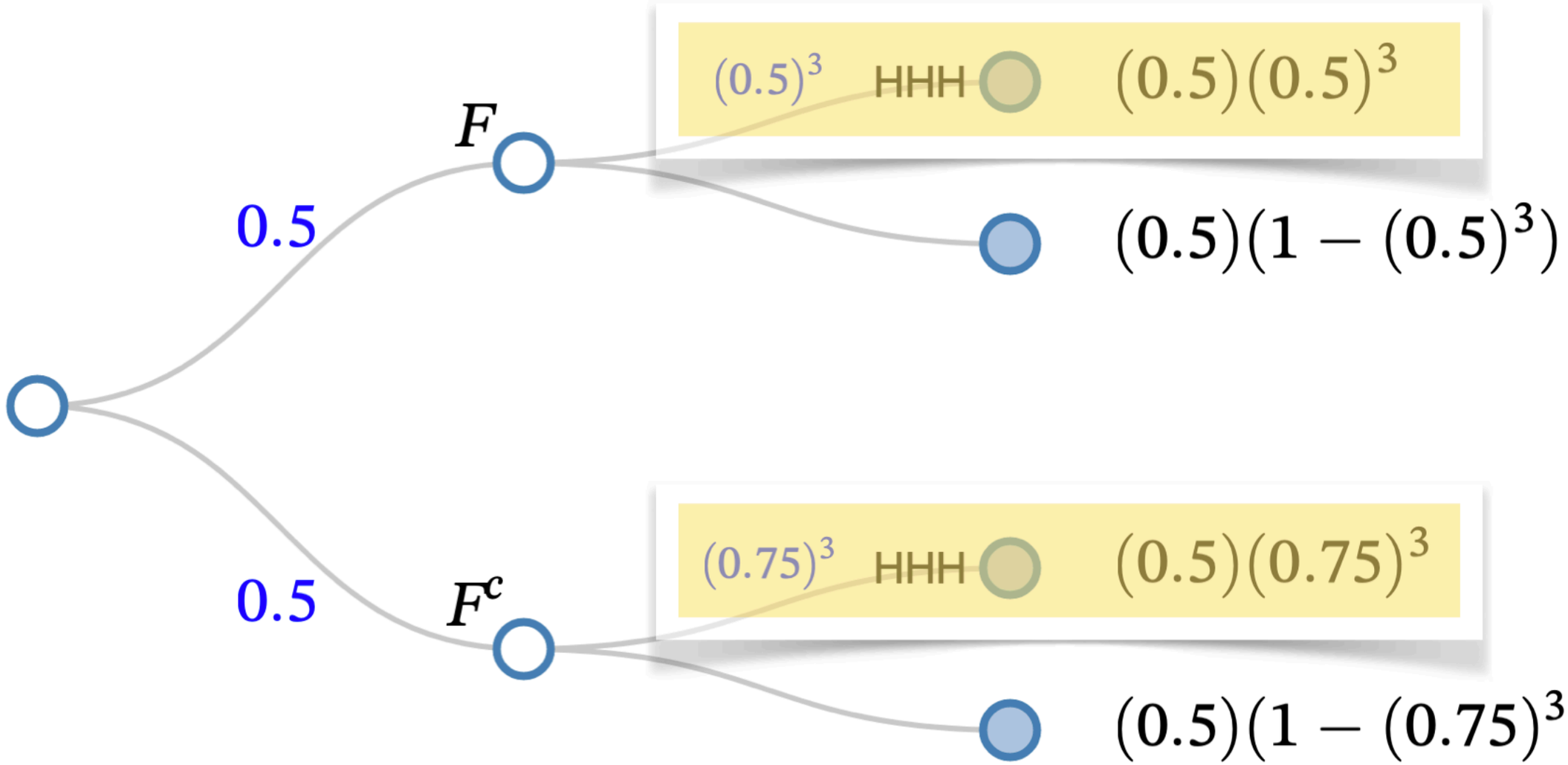
两枚硬币，一枚均匀，另一枚正面概率为3/4。现随机取一枚硬币，连续投掷3次都是正面！取到的是均匀硬币的概率多大?

$$A = \{出现三次正面\}$$

$$F = \{取到均匀硬币\}$$

$$P(F \mid A) = ?$$

$$P(F \mid A) = \frac{P(A \mid F)P(F)}{P(A \mid F)P(F) + P(A \mid F^c)P(F^c)}$$



$$P(F \mid A) = \frac{(0.5)(0.5)^3}{(0.5)(0.5)^3 + (0.5)(0.75)^3}$$

$$P(F \mid A) = \frac{P(A \mid F)P(F)}{P(A \mid F)P(F) + P(A \mid F^c)P(F^c)}$$

Example: medical diagnosis

As an example of how to use this rule, consider the following medical diagonsis problem. Suppose you are a woman in your 40s, and you decide to have a medical test for breast cancer called a **mammogram**. If the test is positive, what is the probability you have cancer? That obviously depends on how reliable the test is. Suppose you are told the test has a **sensitivity** of 80%, which means, if you have cancer, the test will be positive with probability 0.8. In other words,

$$p(x = 1|y = 1) = 0.8 \tag{2.8}$$

where $x = 1$ is the event the mammogram is positive, and $y = 1$ is the event you have breast cancer. Many people conclude they are therefore 80% likely to have cancer. But this is false! It ignores the prior probability of having breast cancer, which fortunately is quite low:

$$p(y = 1) = 0.004 \tag{2.9}$$

Ignoring this prior is called the **base rate fallacy**. We also need to take into account the fact that the test may be a **false positive** or **false alarm**. Unfortunately, such false positives are quite likely (with current screening technology):

$$p(x = 1|y = 0) = 0.1 \tag{2.10}$$

Combining these three terms using Bayes rule, we can compute the correct answer as follows:

$$p(y = 1|x = 1) = \frac{p(x = 1|y = 1)p(y = 1)}{p(x = 1|y = 1)p(y = 1) + p(x = 1|y = 0)p(y = 0)} \tag{2.11}$$

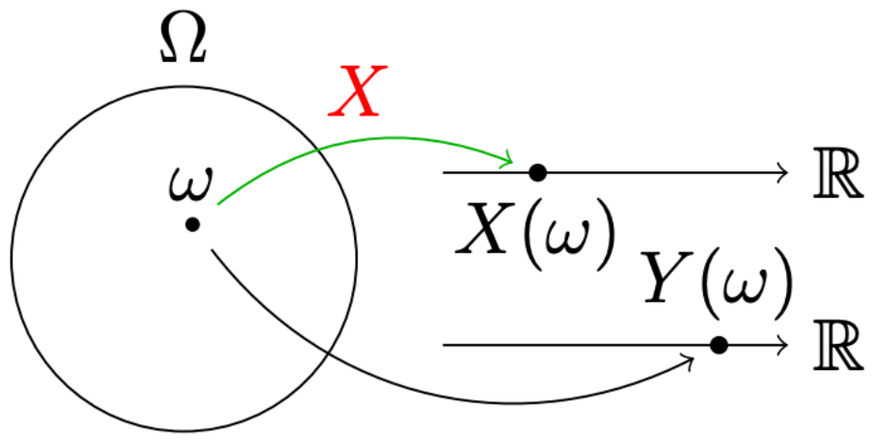
$$= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031 \tag{2.12}$$

where $p(y = 0) = 1 - p(y = 1) = 0.996$. In other words, if you test positive, you only have about a 3% chance of actually having breast cancer![3]

Random Variables

定义 1.1(随机变量的直观描述)　如果条件 S 实现下的情况可以用一个数量 X 来描述，X 究竟等于多少不能预先确定，而随着条件 S 下的结果不同而可能变化，但对任何实数 c，事件“X 取值不超过 c”是有概率的，则把这样一种变量 X 叫做**随机变量**.

定义 1.1′(随机变量的数学描述)　如果条件 S 下所有可能的结果组成集合 $\Omega=\{\omega\}$，$X=X(\omega)$ 是 Ω 上有定义的实值函数，而且对任何实数 c，事件 $\{\omega: X(\omega)\leqslant c\}$ 是有概率的，则称 X 是**随机变量**.(更完全的定义见本章 § 2.4.)



Conditional Independence

We say X and Y are **unconditionally independent** or **marginally independent**, denoted $X \perp Y$, if we can represent the joint as the product of the two marginals (see Figure 2.2), i.e.,

$$X \perp Y \iff p(X, Y) = p(X)p(Y) \tag{2.14}$$

Unfortunately, unconditional independence is rare, because most variables can influence most other variables. However, usually this influence is mediated via other variables rather than being direct. We therefore say X and Y are **conditionally independent** (CI) given Z iff the conditional joint can be written as a product of conditional marginals:

$$X \perp Y|Z \iff p(X, Y|Z) = p(X|Z)p(Y|Z) \tag{2.15}$$

Conditional Independence

Another characterization of CI is this:

Theorem 2.2.1. *$X \perp Y|Z$ iff there exist function g and h such that*

$$p(x, y|z) = g(x, z)h(y, z)$$

for all x, y, z such that $p(z) > 0$.

prove that $p(x|y, z) = p(x|z)$ when $x \perp y|z$

Mass and density function

$$P(x, y) = P(x|y)P(y): \quad \text{product rule}$$

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}: \quad \text{Bayes theorem}$$

$$P(x) = \sum_{y \in \mathcal{Y}} P(x, y) = \sum_{y \in \mathcal{Y}} P(x|y)P(y)$$

Expectation and variance

$$\mathbb{E}[\mathrm{x}] := \int_{-\infty}^{+\infty} x\,p(x)\,dx : \quad \text{mean value,}$$

$$\sigma_{\mathrm{x}}^2 := \int_{-\infty}^{+\infty} (x - \mathbb{E}[\mathrm{x}])^2\,p(x)\,dx : \quad \text{variance}$$

$$\mathbb{E}[f(\mathrm{x})] := \int_{-\infty}^{+\infty} f(x)\,p(x)dx.$$

$$\mathbb{E}_{\mathrm{x},\mathrm{y}}[f(\mathrm{x},\mathrm{y})] = \mathbb{E}_{\mathrm{x}}\left[\mathbb{E}_{\mathrm{y}|\mathrm{x}}[f(\mathrm{x},\mathrm{y})]\right]$$

Given two random variables x, y, their *covariance* is defined as

$$\mathrm{cov}(\mathrm{x},\mathrm{y}) := \mathbb{E}\left[\big(\mathrm{x} - \mathbb{E}[\mathrm{x}]\big)\big(\mathrm{y} - \mathbb{E}[\mathrm{y}]\big)\right],$$

and their *correlation* as

$$r_{\mathrm{xy}} := \mathbb{E}[\mathrm{xy}] = \mathrm{cov}(\mathrm{x},\mathrm{y}) + \mathbb{E}[\mathrm{x}]\,\mathbb{E}[\mathrm{y}].$$

Expectation and variance

Given two random variables x, y, their *covariance* is defined as

$$\operatorname{cov}(\mathrm{x}, \mathrm{y}) := \mathbb{E}\left[\big(\mathrm{x} - \mathbb{E}[\mathrm{x}]\big)\big(\mathrm{y} - \mathbb{E}[\mathrm{y}]\big)\right], \tag{2.28}$$

and their *correlation* as

$$r_{\mathrm{xy}} := \mathbb{E}[\mathrm{xy}] = \operatorname{cov}(\mathrm{x}, \mathrm{y}) + \mathbb{E}[\mathrm{x}]\,\mathbb{E}[\mathrm{y}]. \tag{2.29}$$

A *random vector* is a collection of random variables, $\mathbf{x} = [\mathrm{x}_1, \ldots, \mathrm{x}_l]^T$, and $p(\mathbf{x})$ is the joint PDF (probability mass for discrete variables),

$$p(\mathbf{x}) = p(\mathrm{x}_1, \ldots, \mathrm{x}_l). \tag{2.30}$$

The *covariance matrix* of a random vector $\mathbf{x}$ is defined as

$$\boxed{\operatorname{Cov}(\mathbf{x}) := \mathbb{E}\left[\big(\mathbf{x} - \mathbb{E}[\mathbf{x}]\big)\big(\mathbf{x} - \mathbb{E}[\mathbf{x}]\big)^T\right]: \quad \text{covariance matrix},} \tag{2.31}$$

or

$$\operatorname{Cov}(\mathbf{x}) = \begin{bmatrix} \operatorname{cov}(\mathrm{x}_1, \mathrm{x}_1) & \ldots & \operatorname{cov}(\mathrm{x}_1, \mathrm{x}_l) \\ \vdots & \ddots & \vdots \\ \operatorname{cov}(\mathrm{x}_l, \mathrm{x}_1) & \ldots & \operatorname{cov}(\mathrm{x}_l, \mathrm{x}_l) \end{bmatrix}. \tag{2.32}$$

Expectation and variance

Another symbol that will be used to denote the covariance matrix is $\Sigma_{\mathbf{x}}$. Similarly, the *correlation matrix* of a random vector $\mathbf{x}$ is defined as

$$\boxed{R_{\mathbf{x}} := \mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right]: \quad \text{correlation matrix},} \tag{2.33}$$

or

$$\begin{aligned} R_{\mathbf{x}} &= \begin{bmatrix} \mathbb{E}[\mathrm{x}_1, \mathrm{x}_1] & \dots & \mathbb{E}[\mathrm{x}_1, \mathrm{x}_l] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[\mathrm{x}_l, \mathrm{x}_1] & \dots & \mathbb{E}[\mathrm{x}_l, \mathrm{x}_l] \end{bmatrix} \\ &= \mathrm{Cov}(\mathbf{x}) + \mathbb{E}[\mathbf{x}]\,\mathbb{E}[\mathbf{x}^T]. \end{aligned} \tag{2.34}$$

$$\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\right] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}} \qquad \mathrm{Cov}\left[\mathbf{A}\boldsymbol{x} + \boldsymbol{b}\right] = \mathbf{A}\mathrm{Cov}\left[\boldsymbol{x}\right]\mathbf{A}^{\mathsf{T}}$$

Gaussian distribution

write x $\sim \mathcal{N}(\mu, \sigma^2)$ or $\mathcal{N}(x|\mu, \sigma^2)$, if

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

It can be shown that the corresponding mean and variance are

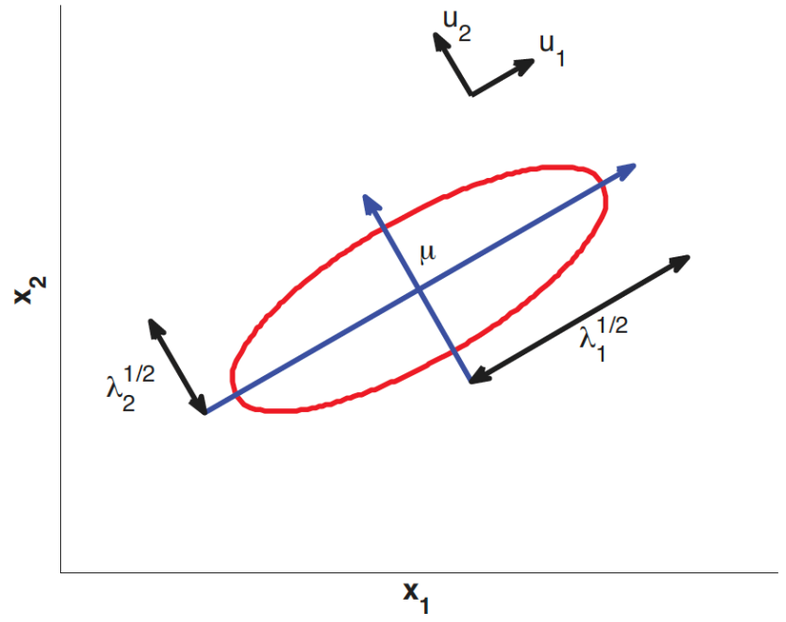
$$\mathbb{E}[\mathrm{x}] = \mu \quad \text{and} \quad \sigma_{\mathrm{x}}^2 = \sigma^2.$$

$$\int_{-\infty}^{+\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = \sqrt{2\pi}\sigma$$

Gaussian distribution

Gaussian or *normal* distribution, $\mathbf{x} \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ with parameters $\boldsymbol{\mu}$ and Σ, which is defined as

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right): \quad \text{Gaussian}$$



Marginal independent if and only it marginals are uncorrelated

Not true in general !

Gaussian distribution

eigendecomposition of $\boldsymbol{\Sigma}$. That is, we write $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, where $\mathbf{U}$ is an orthonormal matrix of eigenvectors satsifying $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues.

Using the eigendecomposition, we have that

$$\boldsymbol{\Sigma}^{-1} = \mathbf{U}^{-T}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^T = \sum_{i=1}^{D} \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^T \tag{4.2}$$

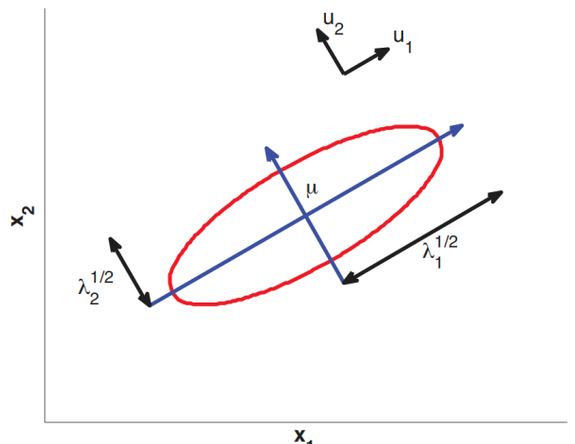
where $\mathbf{u}_i$ is the i'th column of $\mathbf{U}$, containing the i'th eigenvector. Hence we can rewrite the Mahalanobis distance as follows:

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu})^T\left(\sum_{i=1}^{D}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^T\right)(\mathbf{x}-\boldsymbol{\mu}) \tag{4.3}$$

$$= \sum_{i=1}^{D}\frac{1}{\lambda_i}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{u}_i\mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu}) = \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i} \tag{4.4}$$

where $y_i \triangleq \mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu})$. Recall that the equation for an ellipse in 2d is

$$\frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} = 1$$



Parametric modeling - An overview

a *nonnegative* (loss) function,

$$\mathcal{L}(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \longmapsto [0, \infty),$$

and compute $\boldsymbol{\theta}_*$ so as to minimize the total loss, or as we say the *cost*, over all the data points, or

$$f(\cdot) := f_{\boldsymbol{\theta}_*}(\cdot) : \ \boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta} \in \mathcal{A}} J(\boldsymbol{\theta}),$$

where

$$J(\boldsymbol{\theta}) := \sum_{n=1}^{N} \mathcal{L}\big(y_n, f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)\big),$$

In linear regression

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \cdots + \hat{\theta}_l x_l := \hat{\boldsymbol{\theta}}^T \boldsymbol{x}$$

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{N} (y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2$$