Mathematics Notes

Huu Duc Nguyen M.Sc.

27 April 2022

Contents

Abbreviations 1

1 Introduction 2

1.1 Resources . . . 2

2 Matrix 3

2.1 Singular Value Decomposition . . . 3

3 Probabilities 4

3.1 General . . . 4

3.1.1 Basic Definitions . . . 4

3.1.2 Expectation . . . 4

3.1.3 Independence and Variability . . . 5

3.1.4 Bayes Rule . . . 5

3.2 Types of Probability Distributions . . . 5

3.2.1 Bernoulli Distribution . . . 5

3.2.2 Categorical Distribution . . . 6

3.2.3 Univariate Normal Distribution . . . 6

3.2.4 Multivariate Normal Distribution . . . 7

3.2.5 Beta Distribution . . . 7

4 Information Theory 8

4.1 Entropy . . . 8

4.2 Cross Entropy . . . 8

4.3 Kullback - Leiber Divergence . . . 9

4.4 Mutual Information . . . 10

5 Convexity 11

5.1 Convex Sets . . . 11

5.2 Convex function . . . 11

6 Optimization 13

6.1 Convex Optimization Problem . . . 13

6.2 Linear Programming . . . 13

6.3 Quadratic Programming . . . 14

6.4 Geometric Programming . . . 14

7 Search Algorithms **15**
7.1 Graph 15
7.1.1 Undirected Graph 16
7.1.2 Directed Graph 16
7.1.3 Adjacency Matrix 16
7.1.4 Incidence Matrix 16
7.1.5 Trees and Forest 16
7.2 Search Algorithms 16
7.2.1 Properties 16
7.2.2 Depth-first search 16
7.2.3 Breadth-first search 16
7.2.4 Prim's Algorithm 16
7.2.5 Kruskal's Algorithm 16
7.2.6 Dijkstra's Algorithm 16
7.2.7 Bellman and Ford's Algorithm 16
7.2.8 A* Algorithm 16

8 Graphical Models **17**
8.1 Bayesian Network 17

Bibliography **I**

Abbreviations

RL	Reinforcement Learning
prob.	probability
params.	parameters
a.k.a.	also known as
func.	function
pdf.	Probability Density Function
SVD	Singular Value Decomposition
KL	Kullback–Leibler
IG	Information Gain
i.f.f.	if and only if
LP	Linear Programming
QP	Quadratic Programming
TSP	Travelling Salesman Problem

1 Introduction

1.1 Resources

- Essence of linear algebra | 3Blue1Brown
- Essence of calculus | 3Blue1Brown
- Linear Algebra | MIT 18.06 Spring 2005

2 Matrix

2.1 Singular Value Decomposition

Singular Value Decomposition (SVD)

MLcoban.com

3 Probabilities

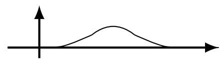
For more examples, exercises with solutions, check:

- Bishop et al. (2006) [BN06].Pattern Recognition and Machine Learning.

3.1 General

3.1.1 Basic Definitions

- If x is discrete: $\sum_x p(x) = 1$ with $\forall\ 0 \leq p(x) \leq 1$
- If x is continuous: $\int p(x)\,dx = 1 \Rightarrow \exists$ a **Probability Density Function (pdf.)**

  $p(x)$ can take any positive value, as long as $\int p(x)\,dx = 1$  **NOTE:**
  : theoretically $p(x) = 0, \forall x$
- Common types

Joint probability:	$p(x_i, y_i)$	$= p(X = x_i, Y = y_i)$
Marginal probability:	$p(x_i)$	$= p(X = x_i)$
Conditional probability:	$p(y_i \vert x_i)$	$= p(Y = y_i \vert X = x_i)$

- Sum rule: $\sum$ joint probability (prob.) = marginal prob.
  $\Rightarrow$ Marginalization
  - discrete variable: $p(x) = \sum_y p(x, y)$
  - continuous variable: $p(x) = \int p(x, y) dy$
- Product rule: Product of marginal prob. and conditional prob. = joint prob.

3.1.2 Expectation

$$\text{For variable } x: \quad \mathbb{E}[x] = \sum_x x.p(x) \quad = \int p(x)x\ dx$$

$$\text{For function } f(\cdot): \quad \mathbb{E}[f(x)] = \sum_x f(x).p(x) \quad = \int p(x)f(x)\ \mathrm{d}x$$

3.1.3 Independence and Variability

- Independence. E.g.: x, y are independent, then

$$\begin{cases} p(x|y) = p(x) \\ p(y|x) = p(y) \end{cases} \Leftrightarrow p(x,y) = p(x).p(y)$$

- Variability
  - variance: how much variability there is in $f(x)$ around its mean value $\mathbb{E}\left[f(x)\right]$
    $var[f] = \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])^2\right] = \mathbb{E}\left[f(x)^2\right] - \mathbb{E}\left[f(x)\right]^2$
  - covariance: for two random variables x, y
    $cov[x,y] = \mathbb{E}_{x,y}\left[xy\right] - \mathbb{E}\left[x\right].\mathbb{E}\left[y\right]$
  - covariance matrix: if x, y are vectors

$$\begin{aligned} cov[\mathbf{x},\mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}\left[\mathbf{x}\right]\}\left\{\mathbf{y}^T - \mathbb{E}\left[\mathbf{y}^T\right]\right\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\mathbf{x}\mathbf{y}^T\right] - \mathbb{E}\left[\mathbf{x}\right].\mathbb{E}\left[\mathbf{y}^T\right] \end{aligned}$$

3.1.4 Bayes Rule

$$\begin{aligned} & p(x_i|y_i).p(y_i) = p(y_i|x_i).p(x_i) = p(x_i, y_i) \\ \Rightarrow\ & p(y_i|x_i) = \frac{p(x_i|y_i).p(y_i)}{p(x_i)} = \frac{p(x_i|y_i).p(y_i)}{\sum\limits_y p(x_i|y_i).p(y_i)} \end{aligned}$$

$\Rightarrow$ the ***Bayes equation***: $\boxed{\textbf{posterior} = \dfrac{\textbf{likelihood} \times \textbf{prior}}{\textbf{normalization factor}}}$

3.2 Types of Probability Distributions

Reference source: machinelearningcoban.com.

3.2.1 Bernoulli Distribution

Bernoulli Distribution is a distribution to describe binary discrete variables. It's the case that the variable can only take value in 2 classes $x \in \{0, 1\}$. E.g., the probability of throwing a coin. The Bernoulli distribution is defined with parameter $\lambda \in [0, 1]$:

$$p(x) = \text{Bern}_x[\lambda] = \begin{cases} p(x=1) = \lambda \\ p(x=0) = 1 - \lambda \end{cases} \tag{3.1}$$

In short form, the above equation can be combined into one:

$$p(x) = \lambda^x (1-\lambda)^{(1-x)} \Rightarrow \begin{cases} p(0) = \lambda^0 (1-\lambda)^1 = 1-\lambda \\ p(1) = \lambda^1 (1-\lambda)^0 = \lambda \end{cases} \tag{3.2}$$

3.2.2 Categorical Distribution

Categorical Distribution is the generalization of *Bernoulli Distribution* for K classes of the discrete variable $x \in \{1, 2, \ldots, K\}$. Accordingly, there will be K parameters to describe this pdf.: $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_K]$, with $\lambda_k \geq 0$ and $\sum \lambda_k = 1$. Each λ_k represents the probability to take the output k: $p(x = k) = \lambda_k$. In short: $p(x) = \text{Cat}_x[\lambda]$.

Another common way to represent the output is the one-hot vector, $\mathbf{x} \in \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K\}$ with $\mathbf{e}_k$ is the k-unit vector, which has all 0-element, except the k-element equal to 1. E.g., given 3 classes: $\mathbf{e}_1 = [1, 0, 0]^T, \mathbf{e}_2 = [0, 1, 0]^T, \mathbf{e}_3 = [0, 0, 1]^T$. We will then have:

$$p(\mathbf{x} = \mathbf{e}_k) = \prod_{j=1}^{K} \lambda_j^{x_j} = \lambda_k \tag{3.3}$$

because for $\mathbf{x} = \mathbf{e}_k$, only $x_k = 1$, while $x_j = 0, \forall j \neq k$.

3.2.3 Univariate Normal Distribution

Univariate Normal Distribution is also known as the Gaussian distribution. For single dimension data (in 1D): $x \in (-\infty, \infty)$, the mean $\mu \in \mathbb{R}$, and the variance σ^2 with $\sigma \in \mathbb{R}$.

$$p(x) = \text{Norm}_x \left[\mu, \sigma^2\right] = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}.\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{3.4}$$

NOTE:

- **Marginals prob. of Gaussian are again Gaussian.**
- When estimating the parameters (params.) of a Gaussian, beware the underestimation problem.

$$\mathbb{E}\left[\mu_{ML}\right] = \mu$$

$$\mathbb{E}\left[\sigma^2_{ML}\right] = \left(\frac{N-1}{N}\right)\sigma^2$$

$$\Rightarrow \tilde{\sigma}^2 = \left(\frac{N}{N-1}\right)\sigma^2_{ML} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$

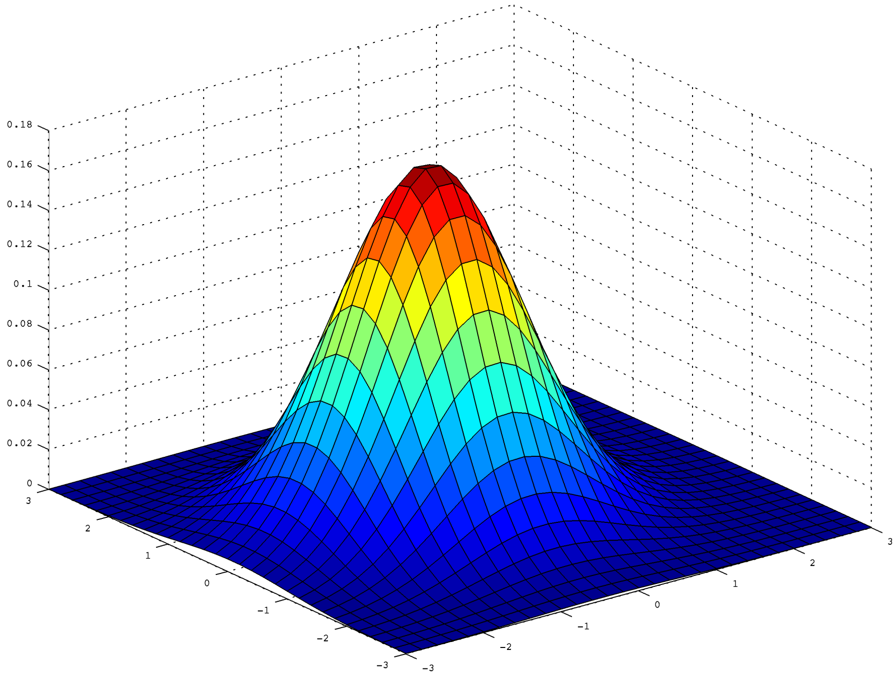


Figure 3.1: Bivariate Gaussian distribution (src).

3.2.4 Multivariate Normal Distribution

Multivariate Normal Distribution is the extension of *Univariate Normal Distribution* to multi-dimensional data: $\mathbf{x}, \boldsymbol{\mu} \in \mathbb{R}^D, \sigma^2 \Rightarrow \Sigma \in \mathbb{S}^D_{++}$ ($\mathbb{S}^D_{++}$ is the set of positive definite symmetric matrix)

$$p(x) = \mathrm{Norm}_x[\boldsymbol{\mu}, \Sigma] = \mathcal{N}(\boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi^{D/2}|\Sigma|^{\frac{1}{2}}}.\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{3.5}$$

3.2.5 Beta Distribution

This distribution describes the parameter for another distributions. E.g., Dirichlet pdf. describes Categorical Distribution (Subsec. 3.2.2)

4 Information Theory

Again, for more examples, exercises with solutions, check:

- Bishop et al. (2006) [BN06].Pattern Recognition and Machine Learning.

4.1 Entropy

$$p(\mathbf{x}) \qquad \text{– distribution (e.g. over observation } \mathbf{x}) \tag{4.1}$$

$$\mathcal{H}(p) = -\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[\log p(\mathbf{x})] \qquad \text{– entropy - how "broad" } p(\mathbf{x}) \text{ is} \tag{4.2}$$

$$= -\sum_{i=1}^{N} p(\mathbf{x}_i) \log p(\mathbf{x}_i) \tag{4.3}$$

Intuition:

- How *random* is the variable?
  The more random the variable, the higher the entropy (Fig. 4.1)
- How large is the log prob. in expectation *under* itself?
  - If you mostly see low log prob., then there are many places with similar prob., and the entropy, as negative log, would be high (Fig. 4.1a).
  - If you mostly see high log prob., then the variable focuses around only a few place, thus, the entropy, as negative log, would be low (Fig. 4.1b).

(a) Example of $p(x)$ with high entropy. **(b)** Example of $p(x)$ with low entropy.

Figure 4.1: Examples of $p(x)$ with different entropy.

4.2 Cross Entropy

The cross entropy between two given prob. distributions p and q is defined as:

$$H(p, q) = \mathbb{E}_p[-\log q] \tag{4.4}$$

With $\mathbf{p}, \mathbf{q}$ as discrete variables:

$$H(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{C} p_i \log q_i \tag{4.5}$$

NOTE: $\nexists \log(0) \Rightarrow$ condition: $\mathbf{q} > 0$

4.3 Kullback - Leiber Divergence

The Kullback–Leibler (KL) divergence, also known as (a.k.a.) *relative entropy*, tells:

- How *different* are two distributions?
- How small is the expected log prob. of one distribution under another, **minus entropy**?
- How many bits of info we expect to loose
- ***A function*** that we can ***optimized***

Example reference: countbayesie

$$H = -\sum_{i=1}^{N} p(x_i) \log p(x_i) \tag{4.6}$$

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) \left[\log p(x_i) - \log q(x_i)\right] \quad \text{discrete case} \tag{4.7}$$

$$= \sum_{i=1}^{N} p(x_i) \log \frac{p(x_i)}{q(x_i)} \tag{4.8}$$

$$= \mathbb{E}_{x \sim p(x)} \left[\log p(x) - \log q(x)\right] \tag{4.9}$$

$$D_{KL}(p||q) = -\int p(\mathbf{x}) \ln q(\mathbf{x}) d\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}\right) \quad \text{continuous case} \tag{4.10}$$

$$= -\int p(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} \tag{4.11}$$

$$D_{KL}(p||q) \geq 0 \tag{4.12}$$

$$D_{KL}(p||q) = 0 \iff p(\mathbf{x}) = q(\mathbf{x}) \tag{4.13}$$

cross entropy = entropy + KL Divergence

$$H(p, q) = H(p) + D_{KL}(p||q) \tag{4.14}$$

E.g.: KL divergence between two normal distributions $\mathcal{N}(\mu_1, \sigma_1)$ and $\mathcal{N}(\mu_2, \sigma_2)$:

$$D_{KL}(p, q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \tag{4.15}$$

PSEUDO CODE: with $\mu_2 = 0, \sigma_2 = 1$

$$\mu, \sigma = \texttt{encoder}(\hat{x}) \tag{4.16}$$

$$\texttt{z} = \mu + \sigma * \texttt{random\_normal(0, 1)} \tag{4.17}$$

$$\texttt{y} = \texttt{decoder(z)} \tag{4.18}$$

$$\texttt{recon\_loss} = \texttt{x.log(y) + (1-x)log(1-y)} \tag{4.19}$$

$$\texttt{KL\_loss} = \frac{1}{2}[\mu^2 + \sigma^2 - \texttt{log}(\sigma^2 + 1e^{-8}) - 1] \tag{4.20}$$

$$\texttt{ELBO} = \texttt{recon\_loss - KL\_loss} \tag{4.21}$$

$$\texttt{loss} = \texttt{-ELBO} \tag{4.22}$$

4.4 Mutual Information

$$p(x) \qquad \text{– distribution (e.g. over observation } \mathbf{x}) \tag{4.23}$$

$$\mathcal{H}(p(\mathbf{x})) = -\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[\log p(\mathbf{x})] \qquad \text{– entropy - how "broad" } p(\mathbf{x}) \text{ is} \tag{4.24}$$

$$\mathcal{I}(\mathbf{x};\mathbf{y}) = D_{KL}\big(p(\mathbf{x},\mathbf{y})||p(\mathbf{x})p(\mathbf{y})\big) \qquad \text{– mutual information} \tag{4.25}$$

$$= \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p(\mathbf{x},\mathbf{y})}\left[\log \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right] \tag{4.26}$$

$$= \mathcal{H}(p(\mathbf{y})) - \mathcal{H}(p(\mathbf{y}|\mathbf{x})) \qquad \text{– relates to IG} \tag{4.27}$$

The last equation implies that we can interpret mutual information $\mathcal{I}(\mathbf{x};\mathbf{y})$ as IG, how much more do we know about $\mathbf{y}$ after receiving observation about $\mathbf{x}$.

E.g.:

- If $\mathbf{x}$ and $\mathbf{y}$ are independent of each others, then $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$, thus $\mathcal{I}(\mathbf{x};\mathbf{y}) = 0$
- If $\mathbf{x}$ and $\mathbf{y}$ depends on each others more and more, the different between the joint prob. $p(\mathbf{x},\mathbf{y})$ and the product of marginal prob. $p(\mathbf{x})p(\mathbf{y})$ grows, thus the $\mathcal{I}(\mathbf{x};\mathbf{y})$ is larger.

5 Convexity

5.1 Convex Sets

- Definition 1: line connects 2 points of a convex set lies within the set
- Definition 2: $\mathcal{C}$ is a convex set if for $\forall x_1, x_2 \in \mathcal{C}$:

$$x_\theta = \theta x_1 + (1-\theta)x_2 \in \mathcal{C}, \quad \forall\, 0 \le \theta \le 1$$

- Hyperplane is a ***convex set***

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = \mathbf{a}^T\mathbf{x} = b, \quad b, a_i \in \mathbb{R}, \quad i \in \mathbb{N}$$

- Half-space is a ***convex set***

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = \mathbf{a}^T\mathbf{x} \le b, \quad b, a_i \in \mathbb{R}, \quad i \in \mathbb{N}$$

- Matrix A is ***positive definite*** if:

$$\mathbf{x}^T\mathbf{A}\mathbf{x} \ge 0, \quad \forall \mathbf{x} \in \mathbb{R}^n \iff \boxed{A \succ 0}$$

  because $\exists A^{-1} \Rightarrow \forall \lambda \ne 0$ (eigenvalues)
- Intersection of convex sets ia a ***convex set***
  $\Rightarrow$ Polyhedra, which is the intersection of halfspaces and hyperplanes ***is convex***.
- x is said to be a ***convex combination*** of $x_1, x_2, \ldots, x_k$ if

$$x = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k \quad \text{with} \quad \theta_1 + \theta_2 + \cdots + \theta_k = 1$$

- ***Convex hull*** of a set $(x_1, x_2, \ldots, x_k)$ is a set of all possible convex combination of that set.
  Convex hull of a set is the smallest convex set that contains that set.
- Two convex sets $\mathcal{C}$ and $\mathcal{D}$ are ***disjoint*** then exist a, b such:

$$\begin{cases} \mathbf{a}^T\mathbf{x} \le b & \forall \mathbf{x} \in \mathcal{C} \\ \mathbf{a}^T\mathbf{x} \ge b & \forall \mathbf{x} \in \mathcal{D} \end{cases}$$

  Set of all $\mathbf{x}$ that $\mathbf{a}^T\mathbf{x} - b = 0$ is a hyperplane that separate $\mathcal{C}$ and $\mathcal{D}$
  $\Rightarrow$ ***separating hyperplane***

5.2 Convex function

Definition: $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function if the domain $\text{dom}(f)$ is a convex set and:

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta)f(y), \quad \forall x, y \in \text{dom}(f), \quad 0 \le \theta \le 1$$

- A function f is ***strictly convex*** if:

  $$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta)f(y)$$

  If there is a minimum point that it is the only minimum point and a global minimum
- ***Affine function:*** $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$ is both convex and concave
  If the variable is a matrix $\mathbf{X}$: $f(\mathbf{X}) = trace(\mathbf{A}^T\mathbf{X}) + b$
- ***Quadratic form:*** $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\ \mathbf{x} + \mathbf{b}^T\mathbf{x} + c$ is
  - convex if $A \succeq 0$
  - concave if $-A \succeq 0$
- A function satisfies 3 norm conditions $\Rightarrow$ convex
- α-subset level of $f : \mathbb{R}^n \to \mathbb{R}$: $\mathcal{C}_\alpha = \{\mathbf{x} \in \mathrm{dom} f | f(\mathbf{x}) \leq \alpha\}$

Checking if f is convex:

- First order condition if and only if (i.f.f.): $\begin{cases} \text{differentiable with convex domain} \\ f(x) \geq f(x_0) + \nabla f(x_0)^T(x - x_0), \quad \forall x, x_0 \in \mathrm{dom} f \end{cases}$
- Second order condition: the Hessian $\nabla^2 f(x) \succeq 0$

6 Optimization

Problem:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f_0(\mathbf{x}) \qquad \text{subject to} \begin{cases} f_i(\mathbf{x}) \leq 0, i = 1, 2, \ldots, m & \text{(inequality constraints)} \\ h_j(\mathbf{x}) = 0, j = 1, 2, \ldots, p & \text{(equality constraints)} \end{cases} \tag{6.1}$$

$$\text{feasible set } \mathcal{D} = \bigcap_{i=0}^{m} \text{dom} f_i \cap \bigcap_{j=0}^{p} \text{dom} h_j \Rightarrow \text{set of all } \mathbf{x} \text{ satisfying all constraints} \tag{6.2}$$

6.1 Convex Optimization Problem

Problem:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f_0(\mathbf{x}) \tag{6.3}$$

$$\text{subject to} \begin{cases} f_i(\mathbf{x}) \leq 0, i = 1, 2, \ldots, m & \text{(inequality constraints)} \\ \mathbf{a}_j^T \mathbf{x} - b_j = 0, j = 1, 2, \ldots, p & \text{(equality constraints)} \\ f_0 \text{ is convex func.} \\ f_i \text{ is convex func.} \\ h_j \text{ is affine func.} \end{cases} \tag{6.4}$$

$$\Rightarrow \begin{cases} f_i(x) \leq 0 \Rightarrow \text{0-sublevel set of } f_i \\ h_j(x) = 0, \quad \forall x \Rightarrow \text{hyperplane} \end{cases} \tag{6.5}$$

⇒ we optimize a convex function in a convex set domain.

6.2 Linear Programming

(Vietnamese: Quy hoach tuyen tinh)

Problem:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \mathbf{c}^T \mathbf{x} + d \qquad \text{subject to} \begin{cases} \mathbf{G}\mathbf{x} \leq \mathbf{h} \\ \mathbf{A}\mathbf{x} = \mathbf{b} \end{cases} \tag{6.6}$$

A standard form of Linear Programming (LP):

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \mathbf{c}^T\mathbf{x} \qquad \text{subject to} \begin{cases} \mathbf{Ax} = \mathbf{b} \\ \mathbf{x} \leq \mathbf{0} \end{cases} \tag{6.7}$$

Python: `cvxopt.solvers.lp`

6.3 Quadratic Programming

Problem:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x} + r \qquad \text{subject to} \begin{cases} \mathbf{Gx} \leq \mathbf{h} \\ \mathbf{Ax} = \mathbf{b} \\ P \succeq 0 \quad (P \text{ is semi-definite}) \end{cases} \tag{6.8}$$

NOTE: When $P = 0$, Quadratic Programming (QP) is LP

Python: `cvxopt.solvers.qp`

6.4 Geometric Programming

- Function $f : \mathbb{R}^n \to \mathbb{R}$ with $\mathrm{dom} f = \mathbb{R}^n_{++}$ (all element > 0) is a ***monomial function*** if:

$$f(x) = c\, x_1^{a_1}\, x_2^{a_2} \dots x_n^{a_n}, \qquad c > 0, \quad a_i \in \mathbb{R} \tag{6.9}$$

- Function f is a ***posynomial function*** if:

$$f(x) = \sum_{k=1}^{K} c_k\, x_1^{a_{1k}}\, x_2^{a_{2k}} \dots x_n^{a_{nk}}, \qquad c_k > 0 \tag{6.10}$$

Problem:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f_0(\mathbf{x}) \qquad \text{subject to} \begin{cases} f_i(\mathbf{x}) \leq 1, \quad i = 1, 2 \dots m \\ h_j(\mathbf{x}) = 1, \quad j = 1, 2 \dots p \\ f_0, f_i \text{ are posynomial func.} \\ h_j \text{ are monomial func.} \end{cases} \tag{6.11}$$

$\Rightarrow$ geometric programming ($x > 0$ is hidden)

Set: $\begin{cases} x_i = e^{y_i} \\ y_i = \log(x_i) \end{cases} \Rightarrow f_0(\mathbf{x}) = \exp(\mathbf{a}^T\mathbf{y} + b)$

Python: `cvxopt.solvers.gp`

7 Search Algorithms

Search problem is one common type of problem which has numerous presences in our lives. The well-known Travelling Salesman Problem (TSP) and its variants are search problems, in which the salesman have to find the shortest route that visit every cities. Many Reinforcement Learning (RL) problems can also be viewed as search problems, in which the machine find the most optimal plan to reach the goal.

A search problem consists of: an agent in a state space, a successor function, a start state and a goal state. The **agent** is the one taking the **action**, e.g., in TSP, the agent is the salesman, and the action is to travel; in RL, the agent is the robot or the machine, and the action could be to move to a different position. The **search state** represents the current situation that the agent is in, which would not necessary equivalent to the world state, which includes every possible details about the environment. E.g., in TSP, the state is the current city, every time the action traveling is taken, the agent moves from one city to another (one state to another). The **successor function** describes the transition from one state to another. This function usually comes with the transition action and costs.

A **solution of search problem** is a sequence of actions (a plan) which transform the start state to a goal state. **Search algorithms** find search solutions, which can be optimal, but in many practical cases, close to optimal within time limitation.

This chapter structures as follows:

- The first section describes how a search problem is formulated mathematically as a graph.
- The second section presents some well-known search algorithms.

7.1 Graph

A graph is the mathematical representation of a search problem. A graph consists of nodes and edges.

7.1.1 Undirected Graph

7.1.2 Directed Graph

7.1.3 Adjacency Matrix

7.1.4 Incidence Matrix

7.1.5 Trees and Forest

7.2 Search Algorithms

7.2.1 Properties

7.2.2 Depth-first search

7.2.3 Breadth-first search

7.2.4 Prim's Algorithm

7.2.5 Kruskal's Algorithm

7.2.6 Dijkstra's Algorithm

7.2.7 Bellman and Ford's Algorithm

7.2.8 A* Algorithm

[**TODO:**]

8 Graphical Models

For more examples, exercises with solutions, check:

- Bishop et al. (2006) [BN06].Pattern Recognition and Machine Learning.

8.1 Bayesian Network

Bayesian networks, a.k.a., Bayes nets, Belief networks and sometimes Causal networks, are graphical representation of a probabilistic model. It comprises of nodes and directed edges.

- Nodes represent variable (continuous or discrete)
- Edges represent dependency between variables

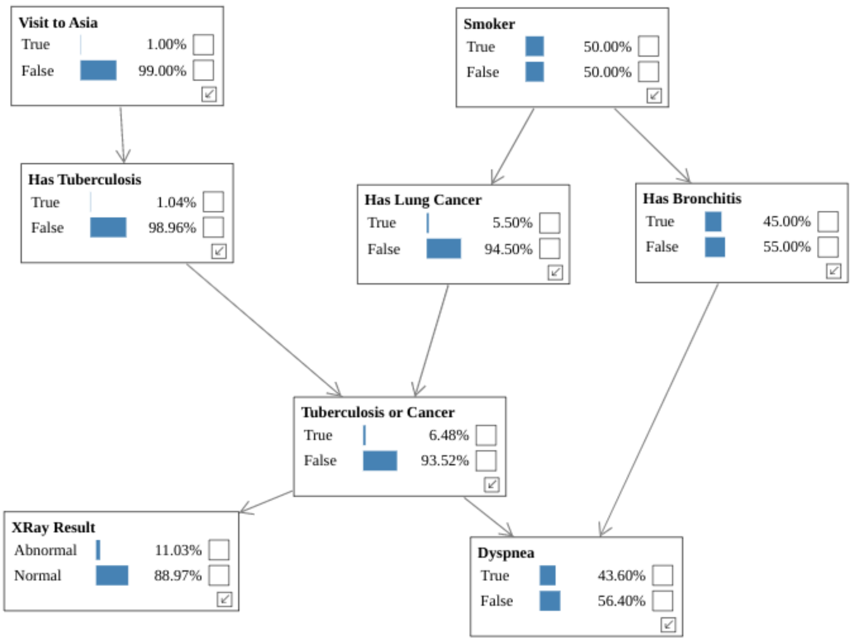


Figure 8.1: Example of Bayesian network: the Asia network (src).

They are useful in a variety of tasks

- Descriptive analytic
- Diagnostic analytic
- Predictive analytic
- Prescriptive analytic

Bibliography

[BN06] C. M. Bishop and N. M. Nasrabadi. *Pattern Recognition and Machine Learning*. Vol. 4. 4. Springer, 2006.