Applied Linear Algebra in Data Analysis

Optimization: A brief introduction

Sivakumar Balasubramanian

Department of Bioengineering
Christian Medical College, Bagayam
Vellore 632002

Optimization

- Optimization is the process of finding the best solution to a problem from a set of possible solutions.
- Optimization problems come up in many applications in engineering, science, economics, biology, medicine, operations research, etc.
- Optimization problems can be classified in different ways, but one major classification gives us: **unconstrained** and **constrained** optimization problems.
- In the context of data analysis, we are often interested in optimization problems of the following form: consider a set of observations $\{\mathbf{a}_i, y_i\}_{i=1}^{m}$. We are interested in identifyin

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \qquad \text{subject to } g_i(\mathbf{x}) \leq 0, \; i = 1, \ldots, m$$

  where, $f(\mathbf{x})$ is the **objective function** and $g_i(\mathbf{x})$ are the **constraint functions**.

Optimization in single variable

- Consider the function $f(x) = x^2 - 2x + 1$.

Fundamental rules of probability

- **Random experiment** – A experiment whose outcome is not predictable.
  - Tossing of a coin.
  - Voltage across a real resistor (R) for a known current.
  - Height and weight of 40 year old person randomly chosen from a population.
- The **outcome** of a random experiment is any observable variable of interest.
- **Sample space** of the experiment S is the universe of possible values we can observe for a random experiment's outcome.
- An **event** of an experiment is any subset of the sample space S.

Fundamental rules of probability

- Consider the experiment tossing a dice, and we observe the count of the dots that turn on the top face of the dice.
  - Observed outcome is an even number. $A = \{2, 4, 6\} \subset S$
  - Observed outcome is a positive number. $A = S \implies$ **Sure event**
  - Observed outcome is 0. $A = \{\} \implies$ **Impossible event**
- For discrete sample spaces and **elementary event** is an event with just single sample point.
- We can combine events to produce other events that might be of interest to us. Set operations can be used to perform algebra on events.

Fundamental rules of probability

- Let A be an event of an experiments, and $p(A)$ the probability of the event A.
- The assignment of probabilities satisfies the following prorperties.
  - For any event A, $0 \leq P(A) \leq 1$.
  - $P(S) = 1$; S is the sample space.
  - For two events A, B,

$$\begin{cases} A \cap B = \emptyset & \implies P(A \cup B) = P(A) + P(B) \\ A \cap B \neq \emptyset & \implies P(A \cup B) = P(A) + P(B) - P(A \cap B) \end{cases}$$

- The other rules for proability calculation for events of an experiment can be derived from these three axioms.
  - $P(\overline{A}) = 1 - P(A)$
  - $A \subset B \implies P(A) \leq P(B)$
  - $P(\emptyset) = 0$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

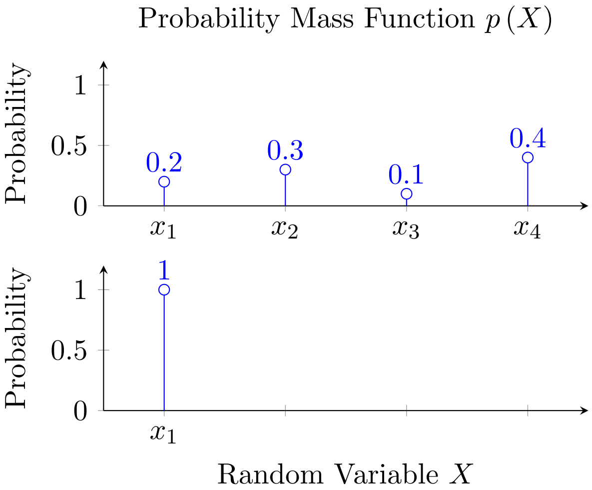
Random variables

- A random variable is X is a function that maps the sample space S to the real numbers $\mathbb{R}$. Random variables allow us to deal with experimental outcomes and event interms of numbers instead of arbitrary symbols. Note: We will use "r.v." to mean "random variable" from this point on.
- Two types of random variables: Discrete random variables and Continuous random variables.
- Discrete random variables take on values from a discrete set of numbers $\mathcal{X}$ (finite or countably infinite).
- Continuous random variables take on values from a continuous set of numbers $\mathcal{X}$ (uncountably infinite).
- Function that assigns probabilities to a discrete random variable X is called the **proability mass function** (p.m.f.) $p\,(X = x)$ is the proability of the random variable X assuming the value x.

$$p\,(X = x) \geq 0, \;\forall x \in \mathcal{X}, \qquad \sum_{\mathcal{X} \in x} p\,(X = x) \geq 0$$

Random variables

Here are two proability mass fucntions.



Joint and Marginal Probabilities

- Consider two r.v. $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$. The joint p.m.f. of these r.v. is defined as,

$$p\left(X = x, Y = y\right) = p\left(\{X = x\} \cap \{Y = y\}\right) = p\left(Y = y, X = x\right)$$

  Meaning of joint probabilities: $p\left(X = x, Y = y\right)$ is the probability of the r.v. X takes on the value x **and** the r.v. Y takes on the value y.
- The marginal p.m.f. of the r.v. X is the probability that it takes on a value x. This can be computed from the joint p.m.f. as the following,

$$p\left(X = x\right) = \sum_{y \in \mathcal{Y}} p\left(X = x, Y = y\right)$$

  Similary the margnal p.m.f. of r.v. Y is

$$p\left(Y = x\right) = \sum_{x \in \mathcal{X}} p\left(X = x, Y = y\right)$$

Conditional probabilities

- Consider two r.v. $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, with the joint p.m.f. $p(X, Y)$.
- The conditional p.m.f $X = x$ given $Y = y$ is defined as,

$$p(X = x | Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}, \text{ if } p(Y = y) \neq 0$$

The conditional proability is not defined if $p(Y = y) = 0$.
Meaning of conditional probabilities: $p(X = x | Y = y)$ is the probability that r.v. X taking on a value $x \in \mathcal{X}$, given that **we know** the r.v. Y has taken on a value $y \in \mathcal{Y}$.
Note that $p(Y = y) = 0$ means that $Y = y$ cannot have occured, so there is nothing to condition on (i.e., the statement "Y has taken on a value $y \in \mathcal{Y}$" is meaningless).

Bayes Rule

Consider two discrete r.v. X and Y. We know the following conditional probabilities,

$$p(X|Y) = \frac{p(X, Y)}{p(Y)} \qquad p(Y|X) = \frac{p(X, Y)}{p(X)}$$

(Note: we drop writing $X = x$ and $Y = y$ for brevity).
Thus, we have the **Bayes rule** or **Bayes theorem**,

$$\begin{aligned} p(X|Y) = \frac{p(Y|X)\,p(X)}{p(Y)} &= \frac{p(Y|X)\,p(X)}{\sum_{x\in\mathcal{X}} p(X = x, Y = y)} \\ &= \frac{p(Y|X)\,p(X)}{\sum_{x\in\mathcal{X}} p(Y|X = x)\,p(X = x)} \end{aligned}$$

Example of applying Bayes rule

You have written a python program that does some clever image processing to automatically detect pulmonary embolism (PB) using a given chest x-ray image. After extensive testing with data from CMC you've estblished that your program has a sensitvity of 85%, i.e. your program will report that a person is +ve for PB from his/her chest x-ray image 85% of the time when the person is indeed +ve for PB. And it has a specificity of 95%, i.e. your program will report that a person is -ve for PB from his/her chest x-ray image 95% of the time when the person is indeed -ve for PB.

When I run your program on my most recent chest x-ray, your program reported that I am +ve for PB! Oh my god! Do I have PB? What is the probability that I have PB?

Independence

We say two r.v. X and Y are unconditionally independent or marginally independent, denoted by $X \perp Y$, if

$$X \perp Y \quad \Longleftrightarrow \quad p(X, Y) = p(X)\, p(Y)$$

What does this mean?

- The two r.v. do not carry any information about the other. Remember the $\perp$ symbol when talking about vectors. $\mathbf{x} \perp \mathbf{y} \implies \mathbf{x}$ is perpendicular to $\mathbf{y}$. Informally, $\mathbf{x}$ does not carry any information about y and *vice versa*. The same idea applies here r.v. X and Y. $X \perp Y \implies$ that r.v. X contains no information about Y and *vice versa*.
- The condition probability is the marginal probability, i.e. $p(X|Y) = p(X)$ and $p(Y|X) = p(Y)$.
- The p.m.f. of X for any given values of Y has the same shape as $p(X)$, and similarly the p.m.f. of Y for any given value of X has the same shape as $p(Y)$.

$$p(X, Y = y) \propto p(X) \qquad p(X = x, Y) \propto p(Y)$$

Conditional Independence

We say two r.v. X and Y are conditionally independent given a r.v. Z, denoted by $X \perp Y|Z$, if

$$X \perp Y|Z \quad \iff \quad p\left(X,Y|Z\right) = p\left(X|Z\right)p\left(Y|Z\right)$$

What does this mean? X carries not information about Y, and *vice versa*, given that we know Z took on some value z.

Theorem: $X \perp Y|Z$ if and only if, there exist functions g and h such that,

$$p\left(X,Y|Z\right) = g\left(X,Z\right)h\left(Y,Z\right)$$

for all X, Y such that $p\left(Z\right) > 0$.

Continuous Random Variables

- Let X be a continuous r.v. such that $X \in \mathcal{X} \subseteq \mathbb{R}$.
- We can meaningfully define probabilities for continuous r.v. only for intervals of the real line. For example, we can define the probability that X takes on a value in the interval $[a, b] \subset \mathcal{X}$.
- For a continuous r.v. X, we define a probability density function (p.d.f.) $f(x)$ such that,

$$p(a \leq X \leq b) = \int_a^b f(X)\, dX$$

  Another useful function is the cummulative distribution function (c.d.f.) $F(X)$, defined as,

$$p(X \leq a) = F(X) = \int_{-\infty}^a f(X)\, dX$$

- For a small interval $[x, x + dx]$, the probability that X takes on a value in this interval is $f(X)\, dx \longrightarrow f(X) = \frac{p(x, x+dx)}{dx}$.

Expected values of a random varaible

Expected value of a r.v is the average value of the r.v. over all possible outcomes. For a discrete r.v. X with p.m.f. $p(X)$, the expected value is,

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot p(X = x)$$

For a continuous r.v. X with p.d.f. $f(X)$, the expected value is,

$$\mathbb{E}[X] = \int_{\mathcal{X}} x \cdot f(X = x)\, dX \quad \text{or} \mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot p(X = x)$$

Expected values of a random varaible

Variance a r.v is a measure of the spread of a r.v. about its mean.

$$\text{var}\left[X\right] = \mathbb{E}\left[\left(X - E\left[X\right]\right)^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2$$

The square root of $\text{var}\left[X\right]$ is called the **standard deviation** of X.

$$\text{std}\left[X\right] = \sqrt{\text{var}\left[X\right]}$$

We can compute the expected value of any function $g\left(\bullet\right)$ of a r.v. X as follows,

$$\mathbb{E}\left[g\left(X\right)\right] = \int_{\mathcal{X}} g\left(X\right) \cdot f\left(X\right) dX$$

Covariance and Correlation between two r.v. X and Y

Consider two r.v. X and Y with joint p.d.f. $f(X, Y)$. The covariance between X and Y measures the (linear) relationship between the two r.v. This is defined as the following,

$$\text{cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

$\text{cov}[X, Y]$ can take on any value between $-\infty$ and ∞.

When $\text{cov}[X, Y]$ is normalized by the standard deviations of X and Y, we get the correlation between X and Y.

$$\text{corr}[X, y] = \frac{\text{cov}[X, Y]}{\sqrt{\text{var}[X]}\sqrt{\text{var}[Y]}}$$

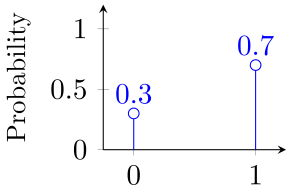
Some discrete r.v. and their p.m.f.

Bernoulli distribution Used to model a single coin toss. The r.v. $X \in \{0, 1\}$ takes on the value 1 if the coin lands heads, and 0 if the coin lands tails. The p.m.f. is,

$$p\left(X = x; \theta\right) = \theta^{X} \cdot (1 - \theta)^{(1-X)}$$

where, p is the probability of the coin turning up heads.

Bernoulli p.m.f. $p\left(X\right)$



Some discrete r.v. and their p.m.f.

Bionomial distribution Used to model the result of experiment with n independent coin tosses. The r.v. $X \in \{0, 1, \ldots, n\}$ takes on the value k if there are k heads in n tosses. The p.m.f. is,

$$p\left(X = k; \theta, n\right) = \frac{n!}{k!\left(n-k\right)!} \cdot \theta^k \cdot \left(1-\theta\right)^{(n-k)}$$

Some discrete r.v. and their p.m.f.

Poisson distribution Used to model the number of events that occur in a fixed interval of time. The r.v. $X \in \{0, 1, \ldots\}$ takes on the value k if there are k events in the interval. The p.m.f. is,

$$p\left(X = k; \lambda\right) = \frac{\lambda^k}{k!} e^{-\lambda}$$

where, λ is the average number of events in the interval.

Some discrete r.v. and their p.m.f.

Uniform distribution Used to model the outcome of an experiment where all outcomes are equally likely. The r.v. $X \in \{a, b\}$ takes on the value x with equal probability. The p.m.f. is,

$$\text{Unif}\left(X = x; a, b\right) = \frac{1}{b - a}\mathbb{I}\left(a \leq x \leq b\right)$$

where, $\mathbb{I}\left(A\right)$ is the indicator function, defined as the following

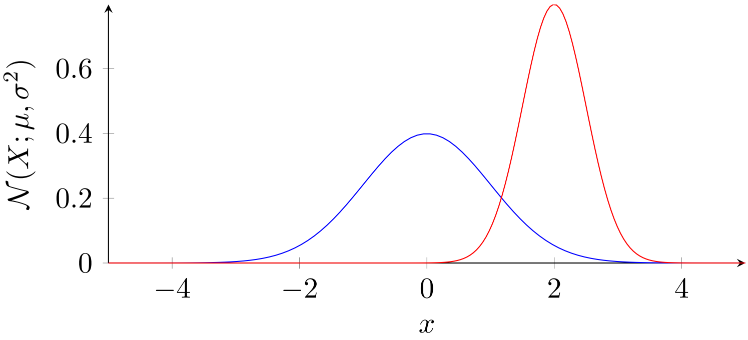
$$\mathbb{I}\left(A\right) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is false} \end{cases}$$

Gaussian (Normal) distribution

Gaussian Distribution is the most commonly used statistical distribution, wose p.m.f. is defined as,

$$\mathcal{N}\left(X = x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where, μ is the mean of the distribution and σ^2 is the variance.



Gaussian (Normal) distribution

- It is commonly observed in nature that many quantities follow a Gaussian distribution.
- Central limit theorem shows that the sum of a large number of independent random variables is approximately Gaussian.
- Its parameters μ and σ^2 have easy interpretations.
- Gaussian distribution is the maximum entropy distribution for a given mean and variance; i.e. it makes the least assumption about the parameter being modelled once we choose the mean and variance.

Multivariate Gaussian (Normal) distribution

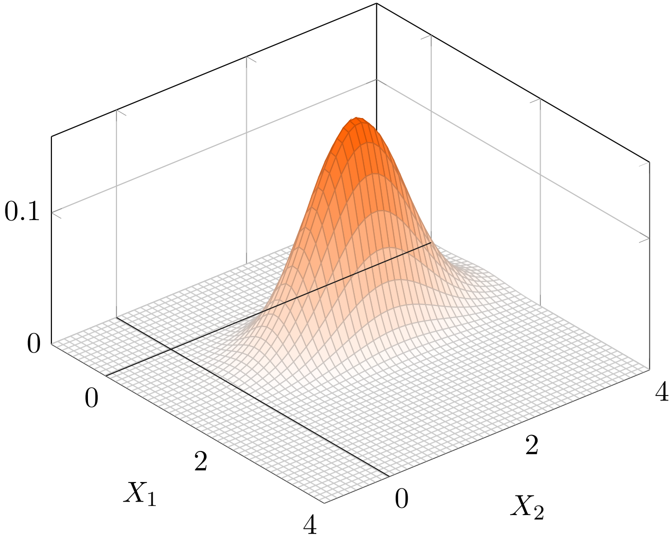
The multivariate Gaussian distribution is commonly use for modelling the joint p.m.f. of multiple r.v.s $X_1, X_2, X_3, \dots X_n$. Let's represent the r.v.s as a vector $\mathbf{x} = \begin{bmatrix} X_1 & X_2 & X_3 & \dots & X_n \end{bmatrix}^\top$. The p.d.f. of the multivariate Gaussian distribution is defined as,

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

where, $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] = \begin{bmatrix} \mathbb{E}[X_1] & \mathbb{E}[X_2] & \cdots \mathbb{E}[X_n] \end{bmatrix}^\top$ is the mean of the distribution, and $\boldsymbol{\Sigma} = \operatorname{cov}[\mathbf{x}]$ is the covariance matrix of the distribution.

$$\begin{aligned} \boldsymbol{\Sigma} =& \operatorname{cov}[\mathbf{x}] = \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right] \\ =& \begin{bmatrix} \operatorname{cov}[X_1, X_1] & \operatorname{cov}[X_1, X_2] & \cdots & \operatorname{cov}[X_1, X_n] \\ \operatorname{cov}[X_2, X_1] & \operatorname{cov}[X_2, X_2] & \cdots & \operatorname{cov}[X_2, X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}[X_n, X_1] & \operatorname{cov}[X_n, X_2] & \cdots & \operatorname{cov}[X_n, X_n] \end{bmatrix} \end{aligned}$$

Multivariate Gaussian Distribution



$$\boldsymbol{\mu} = \begin{bmatrix} 1 & 2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.5 & 0 \\ 0 & 1.0 \end{bmatrix}$$