CSCI567 Machine Learning (Fall 2014)

Drs. Sha & Liu

{feisha,yanliu.cs}@usc.edu

Lecture date: Aug. 27, 2014

Outline

1. Overview
2. Review on Probability
3. Review on Statistics
4. Information Theory
5. Review on Optimization
6. An integrative example

How to grasp machine learning well

Three pillars to machine learning[1]

- Probability, Statistics and Information Theory
- Linear Algebra and Matrix Analysis
- Optimization

Resources to study them

- Suggested Reading:
  - All of Statistics Page 21-89
  - Murphy's textbook
- URL pointers on the syllabus
- Wikipedia (some information might not be 100% accurate, though)

[1] Quote from Prof. Michael I. Jordan

Outline

1. Overview
2. Review on Probability
3. Review on Statistics
4. Information Theory
5. Review on Optimization
6. An integrative example

Probability: basic definitions

Sample Space: a set of all possible outcomes or realizations of some random trial.
Example: Toss a coin twice; the sample space is $\Omega = \{HH, HT, TH, TT\}$.

Event: A subset of sample space
Example: the event that at least one toss is a head is $A = \{HH, HT, TH\}$.

Probability: We assign a real number $P(A)$ to each event A, called the probability of A.

Probability Axioms: The probability P must satisfy three axioms:

1. $P(A) \geq 0$ for every A;
2. $P(\Omega) = 1$;
3. If $A_1, A_2, \ldots$ are disjoint, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Random Variables

Definition: A random variable is a measurable function that maps from a probability space to a measurable space, i.e. $X : \Omega \rightarrow R$, that assigns a real number $X(\omega)$ to each outcome ω.

Example: if $\Omega = \{(x, y) : x^2 + y^2 \leq 1\}$ and our outcomes are samples (x, y) from the unit disk, then these are some examples of random variables: $X(\omega) = x$, $Y(\omega) = y$, $Z(\omega) = x + y$.

Data and Statistics The data are specific realizations of random variables; A statistics is just any function of the data or random variables.

Distribution Function

Definition: Suppose X is a random variable, x is a specific value that it can takes,
Cumulative distribution function (CDF) is the function $F : R \rightarrow [0, 1]$, where $F(x) = P(X \leq x)$.

If X is discrete $\Rightarrow$ *probability mass function*: $f(x) = P(X = x)$.
If X is continuous $\Rightarrow$ *probability density function* for X if there exists a function f such that $f(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f(x)dx = 1$ and for every $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

If $F(x)$ is differentiable everywhere, $f(x) = F'(x)$.

Expectation

Expected Values

- Discrete random variable X, $E[g(X)] = \sum_{x \in \mathcal{X}} g(x) f(x)$;
- Continuous random variable X, $E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)$

Mean and Variance $\mu = E[X]$ is the mean; $var[X] = E[(X-\mu)^2]$ is the variance.
We also have $var[X] = E[X^2] - \mu^2$.

Common Distributions

Discrete variable	Probability function	Mean	Variance
Uniform $X \sim U[1,\ldots,N]$	$1/N$	$\frac{N+1}{2}$	
Binomial $X \sim Bin(n,p)$	$\binom{n}{x} p^x (1-p)^{(n-x)}$	np	
Geometric $X \sim Geom(p)$	$(1-p)^{x-1} p$	$1/p$	
Poisson $X \sim Poisson(\lambda)$	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ	
Continuous variable	Probability density function	Mean	Variance
Uniform $X \sim U(a,b)$	1/ (b-a)	(a + b)/2	
Gaussian $X \sim N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$	μ	
Gamma $X \sim \Gamma(\alpha,\beta)$ $(x \geq 0)$	$\frac{1}{\Gamma(\alpha)\beta^a} x^{a-1} e^{-x/\beta}$	$\alpha\beta$	
Exponential $X \sim exponen(\beta)$	$\frac{1}{\beta} e^{-\frac{x}{\beta}}$	β	

Multivariate Distributions

Definition:

$$F_{X,Y}(x,y) := P(X \leq x, Y \leq y),$$

and

$$f_{X,Y}(x,y) := \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y},$$

Marginal Distribution of X (Discrete case):

$$f_X(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y f_{X,Y}(x,y)$$

or $f_X(x) = \int_y f_{X,Y}(x,y) dy$ for continuous variable.
Conditional probability of X given $Y = y$ is

$$f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Transformation of Random Variables

Let $\mathbb{X} = (X_1, \dots, X_k)$ be a k-dimensional random variable with pdf $f_{\mathbf{X}}(\mathbf{x})$. define a differentiable transformation of $\mathbf{X}$ into $\mathbf{Y}$ using g, such that

$$g(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_k(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} = \mathbf{y}$$

with the inverse $h(\mathbf{y}) = \mathbf{x}$.
The pdf of $\mathbf{Y}$ is $f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{x}}(h(\mathbf{y}))|J(\mathbf{x}, \mathbf{y})|$, where

$$|J(\mathbf{x}, \mathbf{y})| = \text{det}(\begin{bmatrix} \frac{\partial h_1}{\partial y_1} & \dots & \frac{\partial h_1}{\partial y_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_k}{\partial y_1} & \dots & \frac{\partial h_k}{\partial x_k} \end{bmatrix})$$

Example

Suppose X is a random variable, following the *standard normal* distribution

$$X \sim N(0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Then, what is the distribution for $Y = X^2$?

- $X = h_1(Y) = \sqrt{Y}$ or $X = h_2(Y) = -\sqrt{Y}$
- We need to consider each branch, thus

$$f_Y(y) = f_X(h_1(y))|\frac{dh_1(y)}{dy}| + f_X(h_2(y))|\frac{dh_2(y)}{dy}| = \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}}$$

This distribution is called χ^2-distribution.

Bayes Rule

Law of total Probability: X takes values $x_1, \ldots, x_n$ and y is a value of Y, we have

$$f_Y(y) = \sum_j f_{Y|X}(y|x_j) f_X(x_j)$$

Bayes Rule:

(Simple Form)

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

(Discrete Random Variables)

$$f_{X|Y}(x_i|y) = \frac{f_{Y|X}(y|x_i) f_X(x_i)}{\sum_j f_{Y|X}(y|x_j) f_X(x_j)}$$

(Continuous Random Variables)

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{\int_x f_{Y|X}(y|x) f_X(x) dx}$$

Independence

Independent Variables X and Y are *independent* if and only if:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

or $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all values x and y.

IID variables: *Independent and identically distributed* (IID) random variables are drawn from the same distribution and are all mutually independent.

If $X_1, \ldots, X_n$ are independent, we have

$$E[\prod_{i=1}^{n} X_i] = \prod_{i=1}^{n} E[X_i], \quad var[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i^2 var[X_i]$$

Linearity of Expectation: Even if $X_1, \ldots, X_n$ are not independent,

$$E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i].$$

Correlation

Covariance

$$cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)],$$

Correlation coefficients

$$corr(X, Y) = Cov(X, Y)/\sigma_x \sigma_y$$

- Independence $\Rightarrow$ Uncorrelated ($corr(X, Y) = 0$).

However, the reverse is generally not true.
The important special case: multi-variate Gaussian distribution.

Outline

1. Overview
2. Review on Probability
3. Review on Statistics
4. Information Theory
5. Review on Optimization
6. An integrative example

Statistics

Suppose $X_1, \ldots, X_n$ are random variables:

Sample Mean:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Sample Variance:

$$S_{N-1}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2.$$

If X_i are iid:

$$E[\bar{X}] = E[X_i] = \mu,$$
$$Var(\bar{X}) = \sigma^2/N,$$
$$E[S_{N-1}^2] = \sigma^2$$

Point Estimation

Definition The *point estimator* $\hat{\theta}_N$ is a function of samples $X_1, \ldots, X_N$ that approximates a parameter θ of the distribution of X_i.

Sample Bias: The bias of an estimator is

$$bias(\hat{\theta}_N) = E_\theta[\hat{\theta}_N] - \theta$$

An estimator is *unbiased estimator* if $E_\theta[\hat{\theta}_N] = \theta$

Standard error The standard deviation (i.e. the square-root of variance) of $\hat{\theta}_N$ is called the *standard error*

$$se(\hat{\theta}_N) = \sqrt{Var(\hat{\theta}_N)}.$$

Example

Suppose we have observed N realizations of the random variable X:

$$x_1, x_2, \cdots, x_N$$

Then,

- Sample mean $\bar{X} = \frac{1}{N}\sum_n x_n$ is an unbiased estimator of X's mean.
- Sample variance $S^2_{N-1} = \frac{1}{N-1}\sum_n (x_n - \bar{X})^2$ is an unbiased estimator of X's variance
- Sample variance $S^2_N = \frac{1}{N}\sum_n (x_n - \bar{X})^2$ is *not* an unbiased estimator of X's variance

Another example

Suppose we have observed N realizations of the random variable X:

$$x_1, x_2, \cdots, x_N$$

Moreover, suppose we know the true value of X's mean μ. Then,

- Sample variance $S^2_{N-1} = \frac{1}{N-1}\sum_n (x_n - \mu)^2$ is *not* an unbiased estimator of X's variance
- Sample variance $S^2_N = \frac{1}{N}\sum_n (x_n - \mu)^2$ is an unbiased estimator of X's variance

More example

Suppose we have observed N realizations of the random variable X:

$$x_1, x_2, \cdots, x_N$$

Then, in general, neither $\sqrt{S_{N-1}^2}$ nor $\sqrt{S_N^2}$ is an unbiased estimator for σ, i.e., the standard deviation of X.

Outline

1 Overview

2 Review on Probability

3 Review on Statistics

4 Information Theory

5 Review on Optimization

6 An integrative example

Review on Information Theory

Suppose X can have on of the m values: $x_1, \ldots, x_m$. The probability distribution is $P(X = x_i) = p_i$.

Entropy is the smallest possible number of bits, on average, per symbol, needed to transmit a steam of symbols drawn from distribution of X.

$$H(X) = -\sum_{j=1}^{m} p_i \log p_i$$

- "High entropy" means X is from a uniform (boring) distribution;
- "Low entropy" means X is from varied (peaks and valleys) distribution.

Information Theory

Conditional Entropy is the remaining entropy of a random variable Y given that the value of another random variable X is known.

$$H(Y|X) = -\sum_{i=1}^{m} p(X = x_i) H(Y|X = x_i) = -\sum_{i=i}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log p(y_j|x_i)$$

Mutual Information: if Y must be transmitted, how many bits on average would be saved if both ends of the line knew X?

$$I(Y;X) = H(Y) - H(Y|X).$$

Notice that $I(Y;X) = I(X;Y) = H(X) + H(Y) - H(X,Y)$
Kullback-Leibler divergence is a measure of distance between two distributions: a "true" distribution $p(X)$, and an arbitrary distribution $q(X)$.

$$\mathsf{KL}(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

We can write $I(X;Y) = KL(p(x,y)||p(x)p(y))$.

Outline

1. Overview
2. Review on Probability
3. Review on Statistics
4. Information Theory
5. Review on Optimization
6. An integrative example

Optimization

Definition: Optimization refers to choosing the best element from some set of available alternatives. A general form is as follows:

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, i = 1, \ldots, m \\ & h_i(x) = 0, i = 1, \ldots, p. \end{aligned} \tag{1}$$

Difficulties:

1. Local or global optimimum?
2. Difficulty to find a feasible point,
3. Stopping criteria,
4. Poor convergence rate,
5. numerical issues

Convex Optimization

Convex Functions: if for any two points x_1 and x_2 in its domain X and any $t \in [0, 1]$,

$$f(tx_1 + (1-t)x_2) \le tf(x_1) + (1-t)f(x_2).$$

A function f is said to be *concave* if $-f$ is convex.
Convex Set a set S is convex if and only if for any $x_1, x_2 \in S$, $tx_1 + (1-t)x_2 \in S$ for any $t \in [0, 1]$,
Convex Optimization is minimization (maximization) of a convex (concave) function over a convex set.
Examples: Linear Programming (LP), Quadratic Programming (QP), and Semi-Definite Programming (SDP).
Popular convex optimization algorithms:

- Gradient descent
- Conjugate gradient
- Newton's method
- Quasi-Newton method
- Subgradient method

Outline

1. Overview
2. Review on Probability
3. Review on Statistics
4. Information Theory
5. Review on Optimization
6. An integrative example

Outline

Maximum likelihood estimation

Optimization

Convexity

Maximum likelihood estimation (MLE)

Intuitive example

Estimate a coin toss

I have seen 3 flips of heads, 2 flips of tails, what is the chance of head (or tail) of my next flip?

Model

Each flip is a Bernoulli random variable X.

X can take only two values: 1 (head), 0 (tail)

 $p(X=1)=\theta$  $p(X=0)=1-\theta$

Parameter to be identified from data

Principles of MLE

5 (independent) trials

Observations





$X_1 = 1$ $X_2 = 0$ $X_3 = 1$ $X_4 = 1$ $X_5 = 0$

Likelihood of all the 5 observations

$$\theta \quad \times \quad (1-\theta) \quad \times \quad \theta \quad \times \quad \theta \quad \times \quad (1-\theta)$$

$$\mathcal{L} = \theta^3 (1-\theta)^2$$

Intuition

choose θ such that L is maximized

Maximizing the likelihood

Solution

$$\mathcal{L} = \theta^3(1-\theta)^2 \quad \Longrightarrow \quad \theta^{MLE} = \frac{3}{3+2}$$

(Detailed derivation later)

Intuition

Probability of head is the percentage of heads in the total flips.

More generally,

Model (ie, assuming how data is distributed)

$$X \sim P(X; \theta)$$

Training data (observations)

$$\mathcal{D} = \{x_1, x_2, \cdots, x_N\}$$

Maximum likelihood estimate

log-likelihood

$$\mathcal{L}(\mathcal{D}) = \prod_{i=1}^{N} P(x_i; \theta)$$

$$\begin{aligned} \theta^{MLE} &= \arg\max_{\theta} \mathcal{L}(\mathcal{D}) \\ &= \arg\max_{\theta} \sum_{i=1}^{N} \log P(x_i; \theta) \end{aligned}$$

Ex: estimate parameters of Gaussian distribution

Model with unknown parameters

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

Observations

$$\mathcal{D} = \{x_1, x_2, \cdots, x_N\}$$

Log-likelihood

$$\ell(\mu, \sigma) = \sum_{n=1}^{N} \left\{ -\frac{(x_n - \mu)^2}{2\sigma^2} - \log \sqrt{2\pi}\sigma \right\}$$

Solution

We will solve the following later

$$\arg\max_{\mu,\sigma} \ell(\mu, \sigma) = \sum_{n=1}^{N} \left\{ -\frac{(x_n - \mu)^2}{2\sigma^2} - \log \sqrt{2\pi}\sigma \right\}$$

But the solution is given in the below

$$\mu = \bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n \qquad \sigma^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})^2$$

Caveats for complicated models

No closed-form solution

- Use numerical optimization
  - many easy-to-use, robust packages are available
- Stuck in local optimum (more on this later)
  - Restart optimization with random initialization

Computational tractability

- Difficult to compute likelihood $\mathcal{L}(\mathcal{D})$ exactly
- Need to approximate

Optimization

Given an objective function

$$f(x)$$

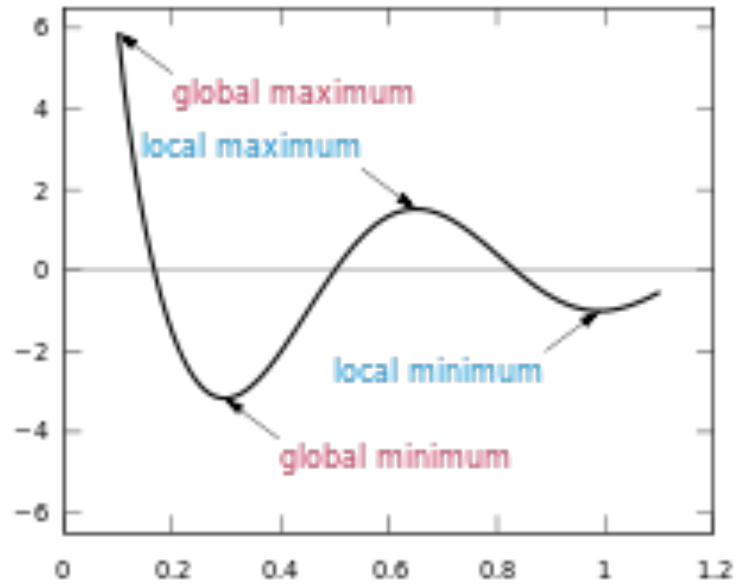
how do we find its minimum

$$\min f(x)$$

optionally, under constraints

$$\text{such that} \quad g(x) = 0$$

difference between global and local optimal

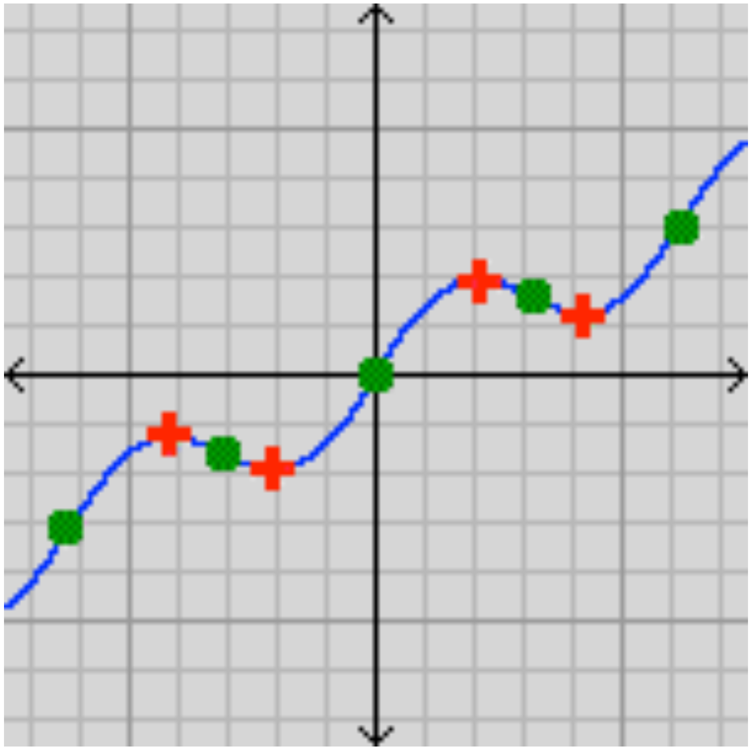


Unconstrained optimization

Fermat's Theorem

Local optima occurs at stationary points, namely, where gradients vanish

$$f'(x) = 0$$



Simple example

What is the minimum of

$$f(x) = x^2$$

Gradient is

$$f'(x) = 2x$$

Set the gradient to zero

$$f'(x) = 0 \rightarrow x = 0$$

Namely, x = 0 is locally optimum (minimum and global, actually)

Remember the MLE of tossing coin?

5 (independent) trials

Observation

$X_1 = 1$ $X_2 = 0$ $X_3 = 1$ $X_4 = 1$ $X_5 = 0$

Likelihood of all the 5 observations

$$\theta \times (1-\theta) \times \theta \times \theta \times (1-\theta)$$

$$\Rightarrow \quad \mathcal{L} = \theta^3 (1-\theta)^2$$

Maximizing the likelihood

the objective function is

$$L(\theta) = \theta^3(1-\theta)^2$$

The gradient is

$$L'(\theta) = 3\theta^2(1-\theta)^2 - 2\theta^3(1-\theta)$$

Set gradient to zero

$$L'(\theta) = 0 \rightarrow \theta = \frac{3}{3+2}$$

Wait a second

The gradient also vanishes if $\theta = 0$

$$L'(\theta) = 3\theta^2(1-\theta)^2 - 2\theta^3(1-\theta)$$

Obviously, $\theta = 0$ does not maximize $L(\theta)$

Thus, be careful

Stationary points are only necessary for (local) optimum

We will discuss sufficient condition later.

Multivariate optimization

Log-likelihood for Gaussian distribution

$$\arg\max_{\mu,\sigma} \ell(\mu,\sigma) = \sum_{n=1}^{N} \left\{ -\frac{(x_n - \mu)^2}{2\sigma^2} - \log\sqrt{2\pi}\sigma \right\}$$

Partial derivatives

$$\frac{\partial \ell}{\partial \mu} = \sum_{n}^{N} -\frac{2(x_n - \mu)}{2\sigma^2}$$

$$\frac{\partial \ell}{\partial \sigma} = \sum_{n}^{N} \left\{ \frac{(x_n - \mu)^2}{\sigma^3} - \frac{1}{\sigma} \right\}$$

Stationary points defined by sets of equations

$$\frac{\partial \ell}{\partial \mu} = 0 \to \mu = \frac{1}{N} \sum_n x_n$$

$$\frac{\partial \ell}{\partial \sigma} = 0 \to \sigma^2 = \frac{1}{N} \sum_n (x_n - \mu)^2$$

We will use the first one to solve the mean

and the second one to compute the standard deviation

a loophole?

In both models, parameters are constrained

- **θ: should be non-negative and be less 1**
- **σ: should be non-negative**

But the optimization did not enforce the constraint

- **yes, we are lucky**

Constrained optimization

General case

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) = 0 \end{aligned}$$

Method of Lagrange multipliers

Construct the following function (Lagrangian)

$$L(x, \lambda) = f(x) + \lambda g(x)$$

Lagrange multiplier

Set derivative to zero

$$\frac{\partial L(x,\lambda)}{\partial x} = f'(x) + \lambda g'(x) = 0$$

Solve x in terms of λ

$$x = h(\lambda)$$

Substitute into constraint, solve λ, then x

$$g(h(\lambda)) = 0$$

Ex: roll a dice

Model

Probability of seeing the number k between 1 and 6

$$P(X = k) = \theta_k$$

Observations

$$\mathcal{D} = \{x_1, x_2, \cdots, x_N\} \qquad x_n \in \{1, 2, \ldots, 6\}$$

Likelihood

$$L(\boldsymbol{\theta}) = \prod_{n=1}^{N} P(X = x_n) = \prod_{k=1}^{6} \theta_k^{n_k}$$

# of times k appear in observations

Optimization

Objective function (log-likelihood)

$$\max \quad \sum_k n_k \log \theta_k$$

constraints

$$\sum_k \theta_k = 1 \qquad\qquad \theta_k \geq 0$$

Lagrangian (ignoring the nonnegative constraint)

$$L(\boldsymbol{\theta}, \lambda) = \sum_k n_k \log \theta_k + \lambda(\sum_k \theta_k - 1)$$

Finding both multiplier and the parameters

Derivatives

$$\frac{\partial L(\boldsymbol{\theta}, \lambda)}{\partial \theta_k} = \frac{n_k}{\theta_k} + \lambda$$

Setting them to zero

$$\theta_k = -\frac{1}{\lambda} n_k$$

Solving the multiplier by using the constraint

$$\sum_k \theta_k = -\frac{1}{\lambda} \sum_k n_k = 1 \rightarrow \lambda = -\sum_k n_k$$

Finally,

$$\theta_k = \frac{n_k}{\sum_k n_k}$$

Intuition: proportional to # of occurrences in observations

Multiple constraints can be handled similarly

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_1(x) = 0 \\ & g_2(x) = 0 \\ & g_3(x) = 0 \end{aligned}$$

Each constraint gets a multiplier

$$L(\boldsymbol{\lambda}, x) = f(x) + \lambda_1 g_1(x) + \lambda_2 g_2(x) + \lambda_3 g_3(x)$$

and use the same stationary point condition

find all multipliers, then the variable x

More difficult situations

Inequality constraints

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq 0 \end{aligned}$$

generally are harder

We won't deal with these types of problems in its most general case

However, we will see some special instances.

Convex optimization

Popular tools in many areas, including machine learning

- **Computationally tractable: as efficient as "linear programming"**
- **Global optimal: no worry of getting not-so-good solutions**

Local vs. global optimal

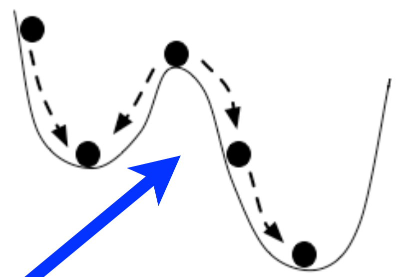
For general objective functions f(x)

We get local optimum

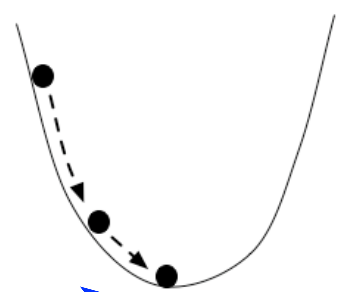
There are special types of functions

where the local optimum is the global optimum

Consider rolling a ball on a hill



depends on where you start



does not depend on where you start

Convex functions

Definition

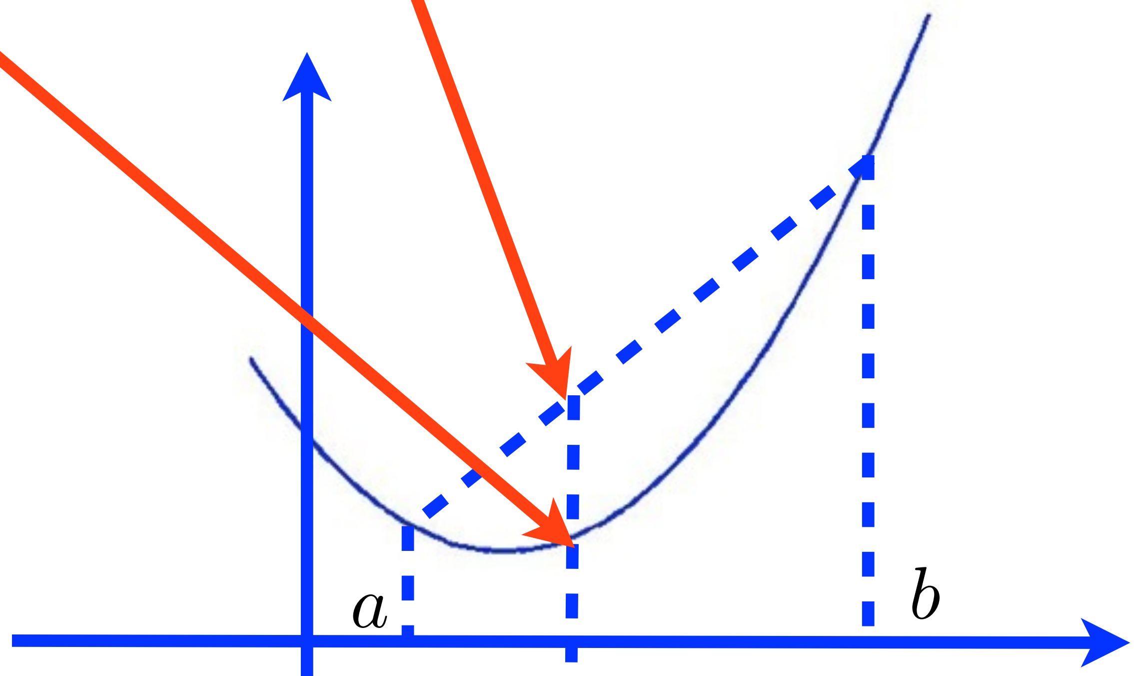
A function f(x) is convex if

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda) f(b)$$

for

$$0 \leq \lambda \leq 1$$

Graphically,



Examples

Convex functions

$$f(x) = x$$

$$f(x) = x^2$$

$$f(x) = e^x$$

$$f(x) = \frac{1}{x} \quad \text{when} \quad x \geq 0$$

Examples

Nonconvex function

$$f(x) = \cos(x)$$

$$f(x) = e^x - x^2$$

Difference in convex functions is not convex

$$f(x) = \log x$$

log (x) is called concave as its negation is convex

How to determine convexity?

f(x) is convex if

$$f''(x) \geq 0$$

Examples

$$(-\log(x))'' = \frac{1}{x^2}$$

We will in future lecture exploit this property

$$(\log(1+e^x))'' = \left(\frac{e^x}{1+e^x}\right)' = \frac{e^x}{(1+e^x)^2}$$

Multivariate functions

Definition

$f(\boldsymbol{x})$ **is convex if**

$$f(\lambda \boldsymbol{a} + (1-\lambda)\boldsymbol{b}) \leq \lambda f(\boldsymbol{a}) + (1-\lambda) f(\boldsymbol{b})$$

How to determine convexity in this case?

Second-order derivative becomes Hessian matrix

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1^2} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_D} \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_2} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_D} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_D} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_D} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_D^2} \end{bmatrix}$$

Convexity for multivariate function

If the Hessian is positive semidefinite, then the function is convex

Ex: $$f(\boldsymbol{x}) = \frac{x_1^2}{x_2}$$

$$\boldsymbol{H} = \begin{bmatrix} \frac{2}{x_2} & -\frac{2x_1}{x_2^2} \\ -\frac{2x_1}{x_2^2} & \frac{2x_1^2}{x_2^3} \end{bmatrix} = \frac{2}{x_2^3} \begin{bmatrix} x_2^2 & -x_1 x_2 \\ -x_1 x_2 & x_1^2 \end{bmatrix}$$

Verify that the Hessian is positive definite

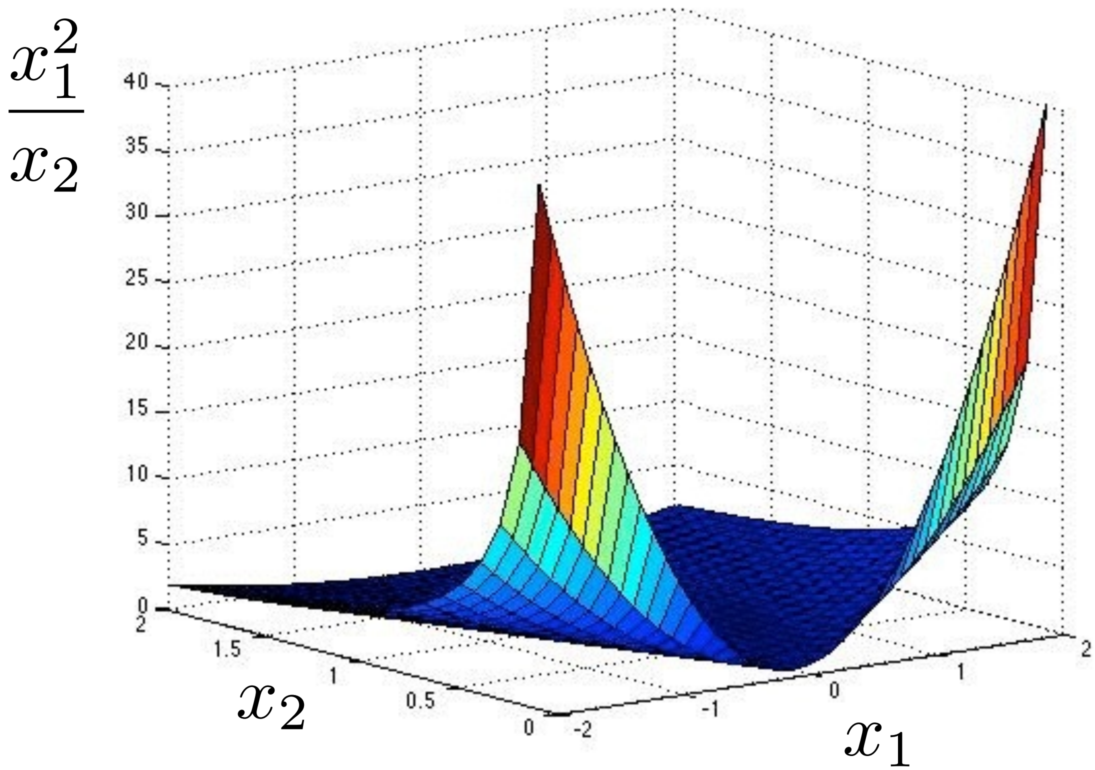
Assume x2 is positive, then

For any vector

$$\boldsymbol{v} = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\begin{aligned}
\boldsymbol{v}^{\mathrm{T}} \boldsymbol{H} \boldsymbol{v} &= \boldsymbol{v}^{\mathrm{T}} \frac{2}{x_2^3} \begin{bmatrix} x_2^2 & -x_1 x_2 \\ -x_1 x_2 & x_1^2 \end{bmatrix} \boldsymbol{v} \\
&= \frac{2}{x_2^3} (a^2 x_2^2 - 2ab x_1 x_2 + b^2 x_1^2) \\
&= \frac{2}{x_2^3} (a x_2 - b x_1)^2 \geq 0
\end{aligned}$$

What does this function look like?



Slightly complicated example

Take-home exercise

Verify the following function

$$f(\boldsymbol{w}) = \log\left(1 + e^{\sum_d w_d x_d}\right)$$

is convex in

$$\boldsymbol{w} = (w_1, w_2, \ldots w_D)^{\mathrm{T}}$$

Why convex function?

if f(x) is convex

then the local optimal

$$\min \quad f(\boldsymbol{x})$$

is also global optimal

This generalizes to constrained optimization

if the constraint

$$g(\boldsymbol{x}) \leq 0$$

define a convex set, namely, for $0 \leq \lambda \leq 1$

$$g(\boldsymbol{a}) \leq 0, g(\boldsymbol{b}) \leq 0 \rightarrow g(\lambda \boldsymbol{a} + (1-\lambda)\boldsymbol{b}) \leq 0$$

Convex set

Take-home exercise

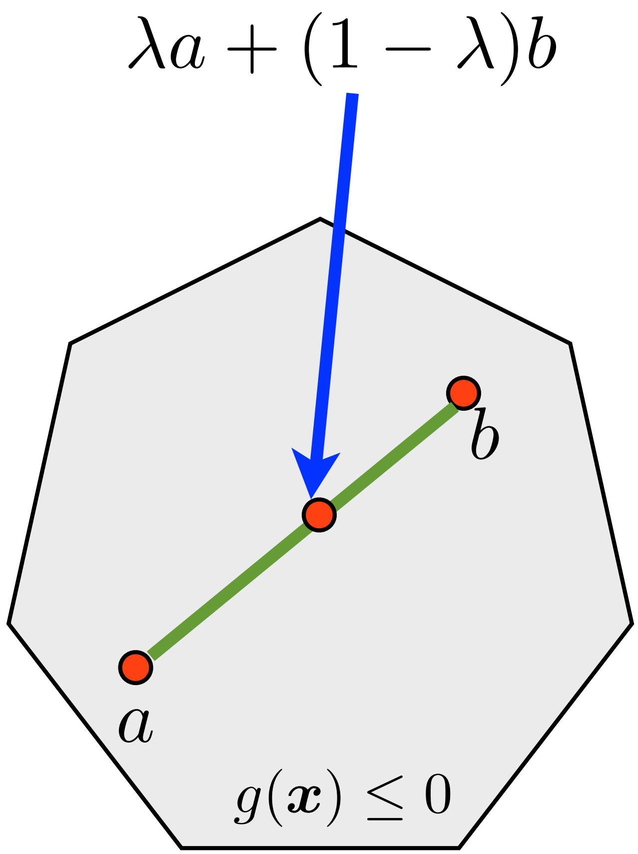
If g(x) is convex

then

$$g(\boldsymbol{x}) \leq 0$$

defines a convex set

graphically,



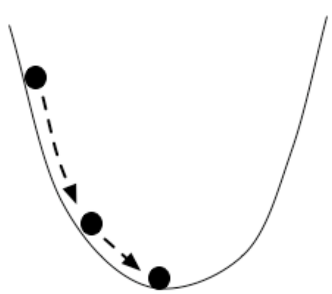
Local vs. global optimal

In practice, convexity can be a very nice thing

In general, convex problems -- minimizing a convex function over a convex set -- can be solved numerically very efficiently

This is advantageous especially if stationary points cannot be found analytically in closed-form

Convex: unique global optimum



nonconvex: local optimum