handsOn03_convex-optimization-probability

September 21, 2016

1 INSTRUCTIONS:

- Take a tag, look at seating chart
- Go to the *seat you were assigned*
- Place the tag on your desk - **we are checking!**

2 EECS 445: Introduction to Machine Learning

2.1 Hands-On Lecture 3: Convex Optimization and Probability

Monday, September 19, 2016

2.2 Outline

This hands-on lecture corresponds to material from **Lecture 03: Convex Optimization & Probability**.

Hands-on Exercises * **Problem 1:** Convex Functions * **Problem 2:** Lagrange Duality * **Problem 3:** Conditional Probability * **Problem 4:** Expectations * **Problem 5:** Variances * **Problem 6:** Maximize Likelihood

2.2.1 Problem 1: Convex Functions

(a) Prove $\forall a, b \in \mathbb{R}$, $f(x) = ax + b$ is convex.

(b) Prove $\forall a \geq 0, \forall b, c \in \mathbb{R}$, $f(x) = ax^2 + bx + c$ is convex.

Hint: Use convex definition: * We say that a function f is *convex* if, for any distinct pair of points x_1, x_2 we have

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2) \quad \forall t \in [0,1]$$

2.2.2 Problem 1: Convex Functions

(c) Prove the following function is convex

$$f(x_1, x_2) = x_1^2 + x_2^2 + x_1 x_2 = \mathbf{x}^T \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix} \mathbf{x}, \forall x_1, x_2 \in \mathbb{R}$$

Hint: Use convex definition: * If f is differentiable, then f is convex iff $f(x+y) \geq f(x) + \nabla_x f(x) \cdot y \quad \forall x, y$

- If f is twice-differentiable, then f is convex iff its hessian is always positive semi-definite!

2.2.3 Solution 1: Convex Functions

(c)

$$\begin{aligned} & f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x}) - \nabla_x f(\mathbf{x})^T \cdot \mathbf{y} \\ = \; & (x_1+y_1)^2 + (x_2+y_2)^2 + (x_1+y_1)(x_2+y_2) - \\ & (x_1^2 + x_2^2 + x_1 x_2) - ((2x_1+x_2)y_1 + (x_1+2x_2)y_2) \\ = \; & y_1^2 + y_2^2 \geq 0 \end{aligned}$$

The Hessian is $\mathbf{H} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, $\forall \mathbf{x} \in \mathbb{R}^2, \mathbf{x}^T H \mathbf{x} = 2x_1^2 + 2x_2^2 + 2x_1x_2 = x_1^2 + x_2^2 + (x_1+x_2)^2 \geq 0$

2.2.4 Problem 2: Convex Optimization

$$\begin{array}{ll} \text{maximize} & f(x) = x_1 + x_2 \\ \text{subject to} & 4x_1 - x_2 \leq 8 \\ & 2x_1 + x_2 \leq 10 \\ & 5x_1 - x_2 \geq -2 \end{array}$$

(a) Convert the above linear programming problem into the standard form.

Hint: Standard form of primal problem: objective function should be **minimize**, constraints only contain "$\leq$" and "$=$", **i.e.**

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & g_i(\mathbf{x}) \leq 0, \quad i = 1, ..., m \\ & h_j(x) = 0, \quad j = 1, ..., n \end{array}$$

2.2.5 Solution 2: Convex Optimization

(a)

$$\begin{array}{lll} \text{minimize} & f(x) = -x_1 - x_2 & \\ \text{subject to} & 4x_1 - x_2 - 8 & \leq 0 \\ & 2x_1 + x_2 - 10 & \leq 0 \\ & -5x_1 + x_2 - 2 & \leq 0 \end{array}$$

2.2.6 Problem 2: Convex Optimization

(b) Derive the dual problem.

Hint: * Its Lagrangian is $L(x, \boldsymbol{\lambda}, \boldsymbol{\nu}) := f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^n \nu_j h_j(x)$

- The **Langragian dual function** is $L_D(\boldsymbol{\lambda}, \boldsymbol{\nu}) \triangleq \min_x L(x, \boldsymbol{\lambda}, \boldsymbol{\nu})$
  $= \min_x \left[f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^n \nu_j h_j(x) \right]$

2.2.7 Solution 2: Convex Optimization

(b)

$$\begin{aligned} L(x,\boldsymbol{\lambda},\boldsymbol{\nu}) &= f(x)+\textstyle\sum_{i=1}^{m}\lambda_i g_i(x)+\sum_{j=1}^{n}\nu_j h_j(x) \\ &= -x_1-x_2+\lambda_1(4x_1-x_2-8)+ \\ &\quad \lambda_2(2x_1+x_2-10)+\lambda_3(-5x_1+x_2-2) \\ &= x_1(4\lambda_1+2\lambda_2-5\lambda_3-1)+x_2(-2\lambda_1+\lambda_2+ \\ &\quad \lambda_3-1)+(-8\lambda_1-10\lambda_2-2\lambda_3) \end{aligned}$$

Dual problem is

$$\begin{array}{lll} \text{maximize} & L_D(\boldsymbol{\lambda}) & = \min_x L(x,\boldsymbol{\lambda}) \\ \text{subject to} & \lambda_i & \geq 0 \end{array}$$

To get feasible solution, we must have the coefficients of x_1 and x_2 to be zero, otherwise $\min_x L(x,\boldsymbol{\lambda})$ would get negetive infinity:

$$\begin{aligned} 4\lambda_1+2\lambda_2-5\lambda_3-1 &= 0 \\ -2\lambda_1+\lambda_2+\lambda_3-1 &= 0 \end{aligned}$$

then $L_D(\boldsymbol{\lambda})$ becomes $-8\lambda_1-10\lambda_2-2\lambda_3$.

2.2.8 Review: Axioms of Probability

- $0 \leq P(A) \leq 1$
- $P(\Omega) = 1$
- If $A_1, A_2, \cdots, A_k$ are disjoint events, then$P(\bigcup_{i=1}^{k} A_i) = \sum_{i=1}^{k} P(A_i)$.

2.2.9 Conditional Probability

For events $A, B \in \mathcal{F}$ with $P(B) > 0$, we may write the **conditional probability of A given B**:

$$P(A|B) = \frac{P(A\cap B)}{P(B)}$$

2.2.10 Problem 3

Suppose x, y are discrete variables and the probability distribution is as follows:

	$x=0$	$x=1$	$x=2$
$y=0$	0.1	0.2	0.3
$y=1$	0.2	0.1	0.1

(a) Compute $p(x=0)$, $p(y=1)$.

(b) Compute $p(x|y=1)$.

(c) Prove the Bayes' Rule

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$

Where B_i is a partition of Ω(note the bottom is just the law of probability).

2.2.11 Problem 4

Given a 2D Gaussian distribution

$$p(x, y) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(x^2 + xy + 2y^2)\}$$

(a) Compute $p(x = 4, y)$.

(b) Compute $p(x = 4)$.

(c) Using Bayse' Rule, compute $p(y \mid x = 4)$.

2.2.12 Solution 4

(a)

$$p(x = 4, y) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(16 + 4y + 2y^2)\}$$

(b)

$$\begin{aligned}
p(x = 4) &= \int_y p(x = 4, y) \\
&= \int_y \tfrac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2}(16 + 4y + 2y^2)\} \\
&= \int_y \tfrac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2}(14 + 2(y + 1)^2)\} \\
&= \tfrac{1}{\sqrt{2\pi}} \exp(-7) \int_y \tfrac{1}{\sqrt{\pi}} \exp\{-\tfrac{1}{2}(2(y + 1)^2)\} \\
&= \tfrac{1}{\sqrt{2\pi}} \exp(-7)
\end{aligned}$$

Notice that $\frac{1}{\sqrt{\pi}} \exp\{-\frac{1}{2}(2(y + 1)^2)\} = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(y - \mu)^2\}$ is Normal distribution with $\mu = -1$ and $\sigma^2 = \frac{1}{2}$, so the integration is 1.

(c)

$$p(y|x = 4) = \frac{p(x = 4, y)}{p(x = 4)} = \frac{1}{\sqrt{\pi}} \exp\{-\frac{1}{2}(2(y + 1)^2)\}$$

2.2.13 Review: Property of Expectation

- $E[a] = a$ for any constant $a \in \mathbb{R}$.
- $E[af(X)] = aE[f(X)]$ for any constant $a \in \mathbb{R}$.
- $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$.
- For a discrete variable X, $E[1X = k] = P(X = k)$.

2.2.14 Problem 5

(a) We have variance of a distribution: $Var[X] = E[X - E[X]]^2$, prove $Var[X] = E[X^2] - E[X]^2$.

Then prove $Var[af(X)] = a^2 Var[f(X)]$ for any constant $a \in \mathbb{R}$.

(b) We also have multiple random variables X and Y, then the covariance is defined as $Cov[X, Y] = E[(X - EX)(Y - EY)]$, prove $Cov[X, Y] = E[XY] - E[X]E[Y]$.

(c) Prove $Var[X + Y] = Var[X] + Var[Y] + Cov[X, Y]$.

2.2.15 Review: Maximum Likelihood

Suppose we have a set of observed data D and we want to evaluate a parameter setting w:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

$$p(D) = \sum_w p(D|w)p(w)$$

We call $p(D|w)$ as the likelihood function. Then we have $p(w|D) \propto p(D|w)p(w)$. Suppose $p(w)$ is the same for all w, we can only choose w to maximize likelihood $p(D|w)$, which is to maximize the log-likelihood $\log p(D|w)$.

2.2.16 Problem 6: Maximum Likelihood Estimation

We have observed data $x_1, \cdots, x_n$ drawn from Bernoulli distribution:

$$p(x) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}$$

(a) What is the likelihood function based on θ?

(b) What is the log-likelihood function?

(c) Compute estimated θ to maximize the log-likelihood function.

2.2.17 Solution 6: Maximum Likelihood Estimation

(a)

$$\begin{aligned} L(\theta) &= p(D|\theta) = p(x_1, \ldots, x_n|\theta) \\ &= \prod_i p(x_i) = \theta^{\sum \mathbb{1}(x_i=1)}(1-\theta)^{\sum \mathbb{1}(x_i=0)} \\ &= \theta^k(1-\theta)^{n-k} \end{aligned}$$

where k is the number of 1s from the observed data.

(b)

$$\log L(\theta) = k \log(\theta) + (n-k)\log(1-\theta)$$

2.2.18 Solution 6: Maximum Likelihood Estimation

(c) Set the derivative of $log(L(\theta))$ to zero we have

$$\frac{\mathrm{d}log(L(\theta))}{\mathrm{d}\theta} = \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0\frac{k}{\theta} = \frac{n-k}{1-\theta}\theta = \frac{k}{n}$$

.

2.2.19 Extra Problem: Maximum Likelihood Estimation for Gaussian Distribution

We have observed data $x_1, \cdots, x_n$ drawn from Normal distribution:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp{(-\frac{1}{2\sigma^2}(x-\mu)^2)}$$

(a) What is the likelihood function based on μ and σ^2?

(b) What is the log-likelihood function?

(c) Compute estimated parameters μ and σ^2 to maximize the log-likelihood function.