CS54701: Information Retrieval

Review of Probability

26 January 2016

Prof. Chris Clifton

1

Probability and Statistics: Outline

- Probability
  - Basic concepts of probability
  - Conditional probability and Independence
  - Common probability distributions
  - Bayes' Rule
- Statistical Inference
  - Statistical learning
  - Maximum likelihood estimation (MLE)
  - Maximum a posterior (MAP) estimation
- Introduction to optimization

Basic Concepts in Probability

- Experiment: flip a coin twice
- Sample space: all possible outcomes of experiment
  - S={HH, HT, TH, TT}
- Event: a subset of possible outcomes (whole space)
  - A={TT} (all tails); B={HT,TH} (1 head and 1 tail)
- Probability of an event: a number indicates how likely the event is
  - Axiom 1: Nonnegativity: $\Pr(A) \geq 0$ for all A belong to S
  - Axiom 2: Normalization: $\Pr(S) = 1$
  - Axiom 3: Additivity: for every sequence of disjoint events

$$\Pr(\bigcup_i A_i) = \sum_i \Pr(A_i)$$

  - Example: Pr(A) = n(A)/N; n(A) size of A, N size of S

Conditional Probability

- If A and B are events with Pr(A)>0, the conditional probability of B given A is

$$\Pr(B \mid A) = \frac{\Pr(A, B)}{\Pr(A)}$$

  - What is the probability of B happens if we already know A happens

- Example

	Male		Female	
Department	Admitted	Not admitted	Admitted	Not admitted
Dept1	40	360	10	90
Dept2	20	80	40	160

Calculate the probabilities

Pr(Admitted | Dept1)

Pr(Admitted | Dept2)

Pr(Admitted | Dept1, Female)

Pr(Admitted | Dept1, Male)

Independence

- Two events A and B are independent iff
- $Pr(A, B) = Pr(A)Pr(B)$
  - The probability of both A and B happens is: probability of A happens times probability of B happens
  - Two events do not have influence on each other
- Example:

	Male		Female	
Department	Admitted	Not admitted	Admitted	Not admitted
Dept1	40	360	10	90
Dept2	20	80	40	160

  - Pr(admitted, male)=60/800=7.5%
  - Pr(admitted)*Pr(male)=110/800*500/800=8.5%

Not independent

Independence

- Two events A and B are independent iff
  $Pr(A, B) = Pr(A)Pr(B)$

This is equal to

- Two events A and B are independent iff
  $Pr(A|B)=Pr(A)$
- Example

	Male		Female	
Department	Admitted	Not admitted	Admitted	Not admitted
Dept1	40	360	10	90
Dept2	20	80	40	160

Not independent

Pr(admitted | male) = 60/500 = 12%

Pr(admitted) = 110/800 = 13.75%

Conditional Independence

- Events A and B are conditionally independent given C iff

  Pr(A,B|C)=Pr(A|C)Pr(B|C)

  – If we know the outcome of event C, then outcomes of event A and B are independent

- Example

	Male		Female	
Department	Admitted	Not admitted	Admitted	Not admitted
Dept1	40	360	10	90
Dept2	20	80	40	160

Pr(Male, Admitted | Dept1)=40/500=8%

Conditionally independent

Pr(Admitted|Dept1)*Pr(male|Dept1)=50/500*400/500=8%

Common Probability Distribution

- Different types of probability distributions associate uncertain outcomes for different physical phenomena
  - Flip a coin: Bernoulli/Binominal
  - Flip a dice (Write a document with several words): Multinomial
  - Random select a point close to a specific point: Gaussian
- Probability mass/density distribution
  - Define how probable the random outcome is a specific event?

  P(X=x) for x in S

Random outcome/variable (e.g., side of a coin)

Specific data point (e.g, head or tail)

Common Probability Distribution

- Some properties of probability mass/density distribution
  - Expectation: the average value of outcomes

$$E(X)=\int x * P(X=x)dx$$

Example: the average outcome of a dice

1/6*1+1/6*2+1/6*3+1/6*4+1/6*5+1/6*6=21/6=3.5

  - Variance: how diverse are the outcomes (deviation from expectation)

$$V(X)=\int (x-E(X))^2 * P(X=x)dx$$

Example: the average outcome of a coin (1 for head, 0 for tail)

$1/2*(0-1/2)^2 + 1/2*(1-1/2)^2 = 1/4$

Common Probability Distributions

Bernoulli/Binomial

- Model binary outcomes: side of a coin, whether a term appears in a document, whether an email is a spam…
  - Bernoulli: binary outcome (i.e., 0 or 1), with probability p to be 1

$$\Pr(X=x\,|\,p)=p^x(1-p)^{1-x};\ x=0,1;\ 0\le p\le 1$$

Expectation: p　　　　Variance: p(1-p)

  - Binomial: n outcomes of a binary variable, the probability p to be 1, what is the probability of outcome 1 appearing x times

$$\Pr(X=x\,|\,n,p)=\binom{n}{x}p^x(1-p)^{n-x};\ x=0,\ldots,n;\ 0\le p\le 1$$

Expectation: np　　　　Variance: np(1-p)

Common Probability Distribution
Multinomial

- Model multiple outcomes: side of a dice; topic of documents; occurrences of terms appear within a document;
  - Multinomial: n outcomes of a variable with multiple values ($v_1..v_n$), with probability p_1 to be v_1,…, probability p_k to be v_k, what is probability of v_1 appearing x_1 times,… v_k appearing x_k times

$$P(X_1 = x_1, \ldots, X_K = x_K \mid n, p_1, \ldots, p_k)$$

$$= \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_K^{x_K}; \sum_{l=1}^{K} x_l = n; 0 \le p_k \le 1; \sum_{l=1}^{K} p_l = 1$$

Expectation: $E(X_i)=np_i$ Variance: $Var(X_i)=np_i(1-p_i)$

Common Probability Distribution
Multinomial

- Examples:

• Three words in vocabulary (sport, basketball, finance), a multinomial model generate the words by probabilities as (p_s=0.5,p_b=0.4,p_f=0.1) (represented by the first character of each word)

A document generated by this model contains 10 words

Question:

- What is the expectation of occurrences of word "sport"?

  10*0.5=5

- What is the probability of generating 5 "sport", 3 "basketball" and 2 "finance

$$\frac{10!}{5!3!2!} 0.5^5 0.4^3 0.1^{0.2}$$

Does the word order matter here? Bag of words representation…

Common Probability Distribution
Gaussian

- Model continuous distribution: draw data points close to a specific point
  - Gaussian (Normal) distribution: select data points close (measured by σ) to a specific point μ .

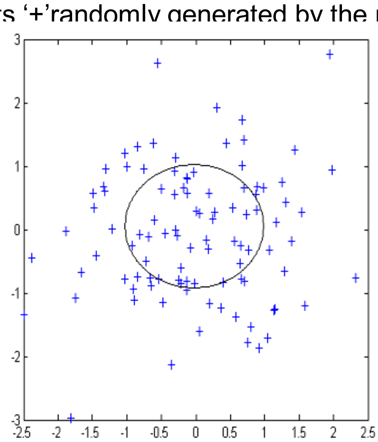
$$\Pr(X = x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2} \right) \quad , \sigma > 0$$

Expectation: E(X_i)= μ Variance: Var(X_i)= σ^2

  - μ, σ^2 can be vectors: multivariate Gaussian

Common Probability Distribution
Gaussian

- Example
  - Gaussian (Normal) distribution with μ=[0 0], σ^2=[1 0;0 1]; 100 data points '+'randomly generated by the model



Bayes's Rule

- Bayes' Rule

Suppose that B_1, B_2, … B_n form a partition of sample space S:

$$B_i \bigcap B_j = \varnothing; \quad \bigcup_i B_i = S$$

Reverse of Conditional Probability Definition

Assume Pr(A) > 0. Then

$$\Pr(B_i \mid A) = \frac{\Pr(A, B_i)}{\Pr(A)} = \frac{\Pr(A \mid B_i)\Pr(B_i)}{\Pr(A)}$$

Additivity

Definition of Conditional Probability

$$= \frac{\Pr(A \mid B_i)\Pr(B_i)}{\sum_{i=1}^{n} \Pr(A, B_i)}$$

$$= \frac{\Pr(A \mid B_i)\Pr(B_i)}{\sum_{i=1}^{n} \Pr(A \mid B_i)\Pr(B_i)}$$

Normalization term

Bayes's Rule

- Interpretation of Bayes' Rule

Hypothesis space: H={H_1, …, H_n}   Observed Data: D

$$P(H_i \mid D) = \frac{P(D \mid H_i)P(H_i)}{P(D)}$$

constant with respect to hypothesis

To pick the most likely hypothesis H*, p(D) can be dropped

Posterior probability of H_i   **Prior probability of H_i**

$$P(H_i \mid D) \propto P(D \mid H_i)P(H_i)$$

Likelihood of data if H_i is true

Common Probability Distribution
Multinomial

- Examples:

  Five words in vocabulary (sport, basketball, ticket, finance, stock)

  Two topics as follows:

  Sport: (p_{sp}=0.4, p_b=0.25, p_t=0.25, p_f=0.1, p_{st}=0)

  Business: (p_{sp}=0.1, p_b=0.1, p_t=0.1, p_f=0.3, p_{st}=0.4)

  Prior Probability: Pr(Sport)=0.5; Pr(Business)=0.5

  Given document $\vec{d}$ =(sport, basketball, ticket, finance)

- What is the probability of $\Pr(\vec{d}|\text{Sport})$, $\Pr(\text{Sport}|\vec{d})$ and $\Pr(\text{Business}|\vec{d})$?
- If we already know Pr(Sport)=0.1; Pr(Business)=0.9, then what about $\Pr(\text{Sport}|\vec{d})$ and $\Pr(\text{Business}|\vec{d})$?

Probability and Statistics: Outline

- Probability
  - Basic concepts of probability
  - Conditional probability and Independence
  - Common probability distributions
  - Bayes' Rule
- Statistical Inference
  - Statistical learning
  - Maximum likelihood estimation (MLE)
  - Maximum posterior (MAP) estimation
- Introduction to optimization

Statistical Inference

- Examples:

  Five words in vocabulary (sport, basketball, ticket, finance, stock)

Two topics “Sport” and “Business”, a set of documents from each topic.
How can we estimate the multinomial distribution for two topics:
e.g., Pr(“sport”|Business), Pr(“stock”|Sport)…

- Probability theory: Model → Data
- Statistical Inference: Data → Model/Parameters
  - Especially with a small amount of observed data
  - In general, statistics has to do with drawing conclusions on whole population based on observations of a sample (data)

Parameter Estimation

- Parameter Estimation:
  - Given a probabilistic model that generates the data in an experiment,
    the model gives a probability of any data $p(D|\theta)$ that depends on the parameter θ
  - We observe some sample data X={x1,…,xn}, what can we say about the value of θ?

Intuitively, take your best guess of θ -- “best” means “best explaining/fitting the data”

Generally an optimization problem

Parameter Estimation

Example:

- Given a document topic model, which is a multinomial distribution

Five words in vocabulary (sport, basketball, ticket, finance, stock)

Observe two documents

$\vec{d}_1$: (sport basketball ticket)

$\vec{d}_2$: (sport basketball sport)

Estimate the parameters of multinomial distribution

$(p_{sp}, p_b, p_t, p_f, p_{st})$

Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation:

- Find model parameters that make generation likelihood reach maximum:

$$M^*=\text{argmax}_M \Pr(D|M)$$

There are K words in vocabulary, $w_1 \dots w_K$ (e.g., 5)

Data: documents $\vec{d}_1,\dots,\vec{d}_I$

For $\vec{d}_i$ with counts $c_i(w_1), \dots, c_i(w_K)$, and length $|\vec{d}_i|$

Model: multinomial M with parameters $\{p(w_k)\}$

Likelihood: $\Pr(\vec{d}_1,\dots,\vec{d}_I|M)$

$$M^*=\text{argmax}_M \Pr(\vec{d}_1,\dots,\vec{d}_I|M)$$

Maximum Likelihood Estimation (MLE)

$$p(\vec{d}_1,..,\vec{d}_I \mid M) = \prod_{i=1}^{I}\left\{\binom{|\vec{d}_i|}{c_i(w_1)\ldots c_i(w_K)}\prod_{k=1}^{K} p_k^{c_i(w_k)}\right\} \propto \prod_{i=1}^{I}\prod_{k} p_k^{c_i(w_k)}$$

$$l(\vec{d}_1,..,\vec{d}_I \mid M) = \log\ p(\vec{d}_1,..,\vec{d}_I \mid M) = \sum_{i=1}^{I}\sum_{k} c_i(w_k)\log\ p_k$$

$$l'(\vec{d}_1,..,\vec{d}_I \mid M) = \sum_{i=1}^{I}\sum_{k} c_i(w_k)\log\ p_k + \lambda(\sum_{k} p_k - 1)$$

Use Lagrange multiplier approach
Set partial derivatives to zero
Get maximum likelihood estimate

$$\frac{\partial l'}{\partial p_k} = \frac{\sum_{i=1}^{I} c_i(w_k)}{p_k} + \lambda = 0 \quad \Rightarrow \quad p_k = -\frac{\sum_{i=1}^{I} c_i(w_k)}{\lambda}$$

$$Since\ \sum_{k} p_k = 1,\ \lambda = -\sum_{k}\sum_{i=1}^{I} c_i(w_k) = -\sum_{i=1}^{I}|\vec{d}_i| \qquad So,\ p_k = p(w_k) = \frac{\sum_{i=1}^{I} c_i(w_k)}{\sum_{i=1}^{I}|\vec{d}_i|}$$

Maximum Likelihood Estimation (MLE)

Example:

- Given a document topic model, what is the multinomial distribution

Five words in vocabulary (sport, basketball, ticket, finance, stock)

Observe two documents

$\vec{d}_1$: (sport basketball ticket)

$\vec{d}_2$: (sport basketball sport)

Maximum likelihood parameters of multinomial distribution

(p_{sp}, p_b, p_t, p_f, p_{st})=(3/6, 2/6, 1/6, 0/6, 0/6)

so (p_{sp}=0.5, p_b=0.33, p_t=0.17, p_f=0, p_{st}=0)

Maximum A Posterior (MAP) Estimation

Maximum Likelihood Estimation:

- Zero probabilities with small sample (e.g., 0 for finance)
- Purely data driven, cannot incorporate prior belief/knowledge

Maximum A Posterior Estimation:

- Select a model that maximizes the probability of model given observed data

$$M^*=\text{argmax}_M \Pr(M|D)=\text{argmax}_M \Pr(D|M)\Pr(M)$$

  - Pr(M): Prior belief/knowledge
  - Use prior Pr(M) to avoid zero probabilities

Maximum A Posterior (MAP) Estimation

There are K words in vocabulary, $w_1 \ldots w_K$ (e.g., 5)

Data: documents $\vec{d}_1,\ldots,\vec{d}_I$

For $\vec{d}_i$ with counts $c_i(w_1)$, …, $c_i(w_k)$, and length $|\vec{d}_i|$

Model: multinomial M with parameters $\{p(w_k)\}$

Posterior: $\Pr(M|\vec{d}_1,\ldots,\vec{d}_I)$

$$M^*=\text{argmax}_M \Pr(M|\vec{d}_1,\ldots,\vec{d}_I)=\text{argmax}_M \Pr(\vec{d}_1,\ldots,\vec{d}_I|M)\Pr(M)$$

Prior Pr(M) is $\Pr(p_1,\ldots p_K)$: Dirichlet Prior

$$Dir(\vec{p}\,|\,\alpha_1,\ldots,\alpha_K)=\frac{\Gamma(\alpha_1+\cdots+\alpha_K)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_k p_k^{\alpha_k-1}$$

Hyper-parameters

Maximum A Posterior (MAP) Estimation

- Dirichlet Prior is the conjugate prior for multinomial distribution
- For the topic model estimation example, MAP estimator is:

Pseudo count

$$p_k = \frac{\sum_{i=1}^{I} c_i(w_k) + (\alpha_k - 1)}{\sum_{i=1}^{I} |\vec{d}_i| + \sum_k (\alpha_k - 1)}$$

$\vec{d}_1$: (sport basketball ticket)

$\vec{d}_2$: (sport basketball sport)

$\alpha_k = 2$ Maximum a posterior parameters of multinomial distribution

(p_{sp}, p_b, p_t, p_f, p_{st})=((3+1)/(6+5), (2+1)/(6+5), (1+1)/(6+5), 1/(6+5), 1/(6+5)

so (p_{sp}=0.364, p_b=0.27, p_t=0.18, p_f=0.091, p_{st}=0.091)

Introduction to Optimization

Optimization

- The mathematical discipline which is concerned with finding the maxima and minima of functions, possibly subject to constraints.

Example we have seen:

$$\vec{p}^{\,*} = \arg\max_{\vec{p}} p(\vec{d}_1, .., \vec{d}_I \mid M) = \prod_{i=1}^{I} \binom{|\vec{d}_i|}{c_i(w_1)...c_i(w_K)} \prod_{k=1}^{K} p_k^{c_i(w_k)}$$

Introduction to Optimization

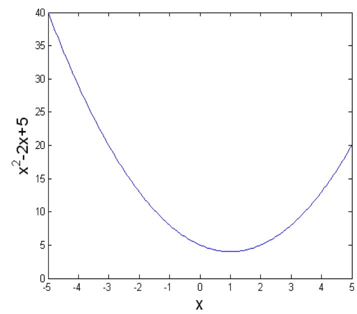
- Calculate analytic solution
  - Calculate the first derivative (with Lagrange multiplier when subjected to constraints)
  - Set the above equation to 0 and try to solve the solution
  - Check whether second derivative is positive (minimum) or negative (maximum)

Example:

$$x^* = \arg\min_x f(x) = \arg\min_x (x^2 - 2x + 5)$$

$$f(x)' = 2x - 2 = 0 \quad \Rightarrow \quad x^* = 1$$

$$f(x^*)'' = 2 > 0 \quad \Rightarrow \quad \textbf{It is minimum}$$

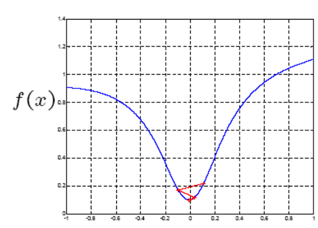


Introduction to Optimization

- Approximate solution with iterative method
  - Many equations by setting derivative to zeros do not have analytic solution
  - Iterative method refines solution step by step
- Newton method uses information of first derivative and second derivative to refine solution

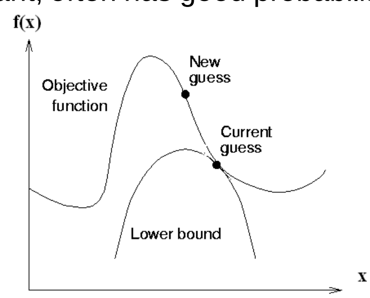
$$x^{(t+1)} = x^{(t)} - \frac{f'(x^{(t)})}{f''(x^{(t)})}$$

New updated solution **Old solution**



Introduction to Optimization

- Newton method does not guarantee improvement of new solution over old one
- Expectation Maximization method
  - Lower bound method, always make improvement
  - More elegant, often has good probabilistic interpretation



Introduction to Optimization

Expectation Maximization method

Examples:

- Given two biased dice A and B with known ($P_A(1)$,…, $P_A(6)$) and ($P_B(1)$,…, $P_B(6)$). Each time, with probability λ draw A, and with probability 1-λ draw B.
- We observe a sequence X={x_1,…, x_n} and want to estimate:

$$\lambda^* = \arg\max_{\lambda} l(X,\lambda)$$

$$\lambda^* = \arg\max_{\lambda} \sum_{i=1}^{n} \left(\log\left(\lambda p_A(x_i) + (1-\lambda) \log(p_B(x_i)) \right) \right)$$

Introduction to Optimization

Previous solution:

$$\lambda^* = \arg\max_\lambda l(X,\lambda) = \arg\max_\lambda \left[l(X,\lambda) - l(X,\lambda^{(t)}) \right]$$

$$= \arg\max_\lambda \sum_{i=1}^{n} \left(\log \left[\frac{\lambda p_A(x_i) + (1-\lambda) p_B(x_i)}{\lambda^{(t)} p_A(x_i) + (1-\lambda^{(t)}) p_B(x_i)} \right] \right)$$

$$= \arg\max_\lambda \sum_{i=1}^{n} \left(\log \left[\frac{\frac{\lambda^{(t)} p_A(x_i)}{\lambda^{(t)} p_A(x_i)} \lambda p_A(x_i) + \frac{(1-\lambda^{(t)}) p_B(x_i)}{(1-\lambda^{(t)}) p_B(x_i)} (1-\lambda) p_B(x_i)}{\lambda^{(t)} p_A(x_i) + (1-\lambda^{(t)}) p_B(x_i)} \right] \right)$$

Set $\lambda^{(t)}$ $F_{Ai} = \frac{\lambda^{(t)} p_A(x_i)}{\lambda^{(t)} p_A(x_i) + (1-\lambda^{(t)}) p_B(x_i)}$ $F_{Bi} = \frac{(1-\lambda^{(t)}) p_B(x_i)}{\lambda^{(t)} p_A(x_i) + (1-\lambda^{(t)}) p_B(x_i)}$ $st. F_{Ai} + F_{Bi} = 1$

$$\geq \sum_{i=1}^{n} \left(F_{Ai} \log\left[\lambda p_A(x_i)\right] + F_{Bi} \log\left[(1-\lambda) p_B(x_i)\right] \right) + Const$$

Convexity of logarithm function, (Jensen Inequality)

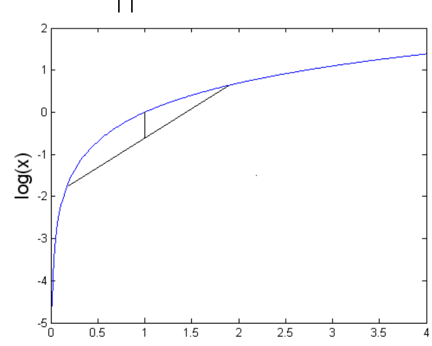
Introduction to Optimization

Set $F_{Ai} = \frac{\lambda^{(t)} p_A(x_i)}{\lambda^{(t)} p_A(x_i) + (1-\lambda^{(t)}) p_B(x_i)}$ $F_{Bi} = \frac{(1-\lambda^{(t)}) p_B(x_i)}{\lambda^{(t)} p_A(x_i) + (1-\lambda^{(t)}) p_B(x_i)}$ $st. F_{Ai} + F_{Bi} = 1$

$$\sum_{i=1}^{n} \left(\log \left[\frac{\frac{\lambda^{(t)} p_A(x_i)}{\lambda^{(t)} p_A(x_i)} \lambda p_A(x_i) + \frac{(1-\lambda^{(t)} p_B(x_i))}{(1-\lambda^{(t)} p_B(x_i))} (1-\lambda) p_B(x_i)}{\lambda^{(t)} p_A(x_i) + (1-\lambda^{(t)}) p_B(x_i)} \right] \right)$$

$$\geq \sum_{i=1}^{n} \left(F_{Ai} \log\left[\lambda p_A(x_i)\right] + F_{Bi} \log\left[(1-\lambda) p_B(x_i)\right] \right) + Const$$

Convexity of logarithm function, (Jensen Inequality)



Introduction to Optimization

Current solution: $\lambda^{(t+1)}$ maximizes derived lower bound

$$\lambda^{(t+1)} = \arg\max_{\lambda} g(\lambda) = \arg\max_{\lambda} \sum_{i=1}^{n} \left(F_{Ai} \log\left[\lambda p_A(x_i)\right] + F_{Bi} \log\left[(1-\lambda) p_B(x_i)\right] \right)$$

$$g(\lambda)' = \sum_{i=1}^{n} \left(\frac{F_{Ai}}{\lambda} - \frac{F_{Bi}}{(1-\lambda)} \right) = 0 \quad \Rightarrow \quad \lambda^{(t+1)} = \frac{\sum_{i=1}^{n} F_{Ai}}{n}$$

Probability and Statistics: Outline

- Probability
  - – Basic concepts of probability
  - – Conditional probability and Independence
  - – Common probability distributions
  - – Bayes' Rule
- Statistical Inference
  - – Statistical learning
  - – Maximum likelihood estimation (MLE)
  - – Maximum posterior (MAP) estimation
- Introduction to optimization

Probability and Statistics: Outline

- References:

Section 2.1, Foundation of Natural Language Processing
http://cognet.mit.edu/library/books/mitpress/0262133601/cache/chap2.pdf (pp. 39-59)

Online notes of probability and Statistics for computer science: http://www.utdallas.edu/~mbaron/3341/Fall06/ (Chaps 2,3,4,12)

Probability and Statistics MH. DeGroot, MJ. Schervish 2001. Addison-Wesley

Optimization (online): “Convex Optimization”, S. Boyd and L. Vandenberghe, http://www.stanford.edu/~boyd/cvxbook/