Notes on Advanced Statistical Machine Learning Course

Dr. Stefanos Zafeiriou

January 2015

1 Introduction

These notes are provided to help students with mathematical derivations, as well as to give a road-map of preliminaries. The course does not have strong prerequisites. Nevertheless, the student should be familiar with or familiarize himself/herself with some basic notions of probability, statistics, linear algebra and basic elements of optimization. A road-map of the basic notions is provided below.

2 Probabilities

- What is a random variable:

  A function which maps events or outcomes to a number set (i.e., integers, real etc).

- What is probability
  - Frequentistic view: Probability of an event is the limit of its relative frequency in a large number of trials.
  - Bayesian view: Probability is a measure of belief regarding the predicted outcome of an event.

- Bayes' theorem (including conditional probabilities)

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

- Probabilities of union and intersection of sets
  - The probability of the union of two (or more) events is the probability that at least one of these events will occur. In the case of two events A and B this probability is denoted by $p(A \cup B)$.
  - The probability of the intersection of two (or more) events is the probability that these events will both occur. In the case of two events A and B this probability is denoted by $p(A \cap B)$.

- Joint and marginal probabilities
  - Joint probability of some random variables is a probability distribution that gives the probability that each of these random variables falls in any particular range or discrete set of values specified for these variables (e.g., $p(x, y)$ is the joint probability for some x, y).
  - Marginal distribution of random variables is the probability distribution of each of the variables independently. In other words, it provides the probabilities of various values of the variables without reference to the values of the other variables (i.e., $p(x), p(y)$ are the marginal distributions of $p(x, y)$).
- Independence and conditional independence
  - Two (or more) random variables are independent if the realization of one does not affect the probability distribution of the other. Two random variables a and b are independent if and only if their joint probability equals the product of their probabilities: $p(a, b) = p(a)p(b)$. Similarly, two sets of events A and B are independent if and only if $p(A \cap B) = p(A)p(B)$.
  - Variables a and b are conditionally independent given c if any of the following holds
    * $\mathrm{p}(a, b|c) = \mathrm{p}(a|c)\mathrm{p}(b|c)$
    * $\mathrm{p}(a, |b, c) = \mathrm{p}(a|c)$
    * $\mathrm{p}(b|a, c) = \mathrm{p}(b|c)$.

    Knowing c contains all the knowledge about b (i.e. a does not contain any information). This is due to the fact that a does not influence b or because c provides all the information that the knowledge about a would provide.
- Probability density function and Cumulative distribution function
  - Probability density function (PDF), of a random variable is a function that describes the relative likelihood for this random variable to take a given value. A function $p(x)$ can be a pdf iff $p(x) \geq 0$ and $\int_{-\infty}^{+\infty} p(x) = 1$. Given the pdf, the probability of $x \in [a, b]$ can be calculated as: $P[a \leq x \leq b] = \int_a^b p(x)\, dx$. For the Gaussian distribution, the pdf is defined as $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(x - \mu)^2\}$.
  - Cumulative distribution function (CDF) is the monotonic function that computes the probability $x \in (-\infty, b]$, or formally $F(b) = \int_{-\infty}^{b} p(x)\, dx$. Similarly, pdf can be computed from the cdf as $p(x) = \frac{dF(x)}{dx}$.
  - Change of variables in pdfs. Assume x_1, x_2 random variables with pdfs $p_1(x_1), p_2(x_2)$. Also assume that $x_2 = g(x_1)$. Then, we can derive p_2

from p_1 and g. Starting from the fact that the probability contained in a differential area must be invariant under change of variables we get

$$\begin{aligned} p(x_2)dx_2 &= p(x_1)dx_1 \\ p(x_2) &= |\frac{dx_1}{dx_2}|p(x_1) \\ p(x_2) &= |\frac{dg^{-1}(x_2)}{dx_2}|p(g^{-1}(x_2)). \end{aligned}$$

3 Statistics

- Expected value and variance. The expected value of a random variable x is defined as
  $E(x) = \int_{-\infty}^{+\infty} xp(x)dx$.
  The variance of a random variable x is defined as $var(x) = E(x^2) - (E(x))^2$. The expected value of a Gaussian random variable x (also written as x follows $N(x|\mu, \sigma^2)$) is $E(x) = \mu$ and its variance is $var(x) = \sigma^2$.

- Maximum likelihood (ML) and maximum a posteriori probability (MAP) estimates. Lets assume a population of n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and θ the set of parameters of the data generation model. Then, the Maximum Likelihood (ML) estimate of optimum θ is given by maximizing the joint likelihood

$$\theta_o = \arg\min_{\theta} p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\theta). \tag{1}$$

  Lets assume again a population of n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and θ a set of parameters of the assumed generative model. Furthermore, let $g(\theta)$ be a prior distribution on θ. The Maximum a-posteriori (MAP) estimate is then given by maximimizing the joint distribution

$$\theta_o = \arg\min_{\theta} p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \theta) = \arg\min_{\theta} p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\theta)p(\theta). \tag{2}$$

  Examples of the above are provided in the course slides.

4 Linear Algebra

- What is a vector, what is a matrix
  - Vector $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = (x_1 \ldots x_n)^T = [x_i]$
  - Matrix $\mathbf{A} = \begin{bmatrix} a_{11} & \ldots & a_{1l} \\ \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{nl} \end{bmatrix} = (\mathbf{a}_1 \ldots \mathbf{a}_l) = \begin{bmatrix} \tilde{\mathbf{a}}_1^T \\ \vdots \\ \tilde{\mathbf{a}}_n^T \end{bmatrix} = [a_{ij}]$
- What is a vector's inner and outer product

  - Inner (dot) product $\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i$
  - Outer product $\mathbf{x}\mathbf{y}^T = \begin{bmatrix} x_1y_1 & \dots & x_1y_n \\ \vdots & \ddots & \vdots \\ x_ny_1 & \dots & x_ny_n \end{bmatrix}$

- Matrix-matrix and matrix-vector multiplication, identity matrix
  - Matrix multiplication $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{n\times l}$, $\mathbf{B} = [b_{ij}] \in \mathbb{R}^{l\times m}$

$$\mathbf{AB} = \left[\sum_{i=1}^{M} a_{ik}b_{kj}\right] = \begin{bmatrix} \tilde{\mathbf{a}}_1^T \\ \vdots \\ \tilde{\mathbf{a}}_n^T \end{bmatrix} [\mathbf{b}_1 \dots \mathbf{b}_m] = [\tilde{\mathbf{a}}^T\mathbf{b}_j]$$

$$= [\mathbf{A}\mathbf{b}_1 \dots \mathbf{A}\mathbf{b}_m] = \begin{bmatrix} \tilde{\mathbf{a}}_1^T\,\mathbf{B} \\ \vdots \\ \tilde{\mathbf{a}}_n^T\,\mathbf{B} \end{bmatrix}$$

  - Matrix-vector multiplication $\mathbf{Ab} = \begin{bmatrix} \tilde{\mathbf{a}}_1^T \\ \vdots \\ \tilde{\mathbf{a}}_n^T \end{bmatrix} \mathbf{b} = \begin{bmatrix} \tilde{\mathbf{a}}_1^T\mathbf{b} \\ \vdots \\ \tilde{\mathbf{a}}_n^T\mathbf{b} \end{bmatrix} = [\tilde{\mathbf{a}}_j^T\mathbf{b}]$
  - Identity matrix $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \; 3 \times 3 \text{ Identity Matrix}$$

- Matrix transposition $\mathbf{A}^T = \begin{bmatrix} a_{11} & \dots & a_{n1n} \\ \vdots & \ddots & \vdots \\ a_{l1} & \dots & a_{nl} \end{bmatrix} = (\tilde{\mathbf{a}}_1 \dots \tilde{\mathbf{a}}_n) = \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_l^T \end{bmatrix}$
- Matrix trace, matrix determinants
  - Matrix trace $\text{tr}(\mathbf{A}) = \sum_{i}^{n} a_{ii}$
  - Matrix determinant $|\mathbf{A}| = \sum_{j=1}^{n} (-1)^{j+k} a_{jk} |\mathbf{A}_{jk}|$
- Matrix inverse and pseudo-inverse
  - Matrix inverse $\mathbf{A^{-1}}$, $\mathbf{A^{-1}A} = \mathbf{AA^{-1}} = \mathbf{I}$, $\mathbf{A^{-1}}$ exists iff $|\mathbf{A}| \neq 0$
  - Matrix pseudo-inverse $\mathbf{A}^+ = \left(\mathbf{A^T A}\right)^{\mathbf{-1}} \mathbf{A^T}$, $\mathbf{A}^+\mathbf{A} = \mathbf{I}$
- Rank of a matrix

  - Definition 1: is the dimension of the largest square sub-matrix of a matrix that has a non-zero determinant.
  - Definition 2: is the maximum number of linearly independent columns (or rows) of a matrix

- Orthogonal/orthonormal matrices
  - Orthogonal $\mathbf{Ab} = \begin{bmatrix} \tilde{\mathbf{a}}_1^T \\ \vdots \\ \tilde{\mathbf{a}}_n^T \end{bmatrix} \tilde{\mathbf{a}}_\mathbf{j}^\mathbf{T}\tilde{\mathbf{a}}_\mathbf{k} = 0$ for every $j \neq k$, $\mathbf{AA^T} = L$
  - Orthonormal $\mathbf{Ab} = [\mathbf{a}_1 \dots \mathbf{a}_n]$ $\mathbf{a_j}^\mathbf{T}\tilde{\mathbf{a}}_\mathbf{k} = 0$ for every $j \neq k$, $\mathbf{AA^T} = \mathbf{I}$
- Matrix Eigenanalysis $\mathbf{Ax} = \lambda\mathbf{x},\ \mathbf{x} \neq \mathbf{0}$
- Matrix Diagonalization: Given a matrix $\mathbf{A}$, find $\mathbf{P}$ such that $\mathbf{P^{-1}AP}$ is diagonal
- Gram-Schmidt orthogonalization
  - A process to convert a non-orthonomal basis $\mathbf{S} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ to an orthonormal one for the inner product space $\mathbf{V}$
  - The key idea of the Gram-Schmidt orthogonalization is to subtract from every new vector, $\mathbf{u}_k$, its components in the directions already determined, $\{\mathbf{v_1}, \mathbf{v_2}, \dots, \mathbf{v_{k-1}}\}$

5 Optimization

During the course we will formulate various optimization problems in which $\mathbf{x}$ is a vector or matrix. The optimization problems that we will encounter are of the form

- without constraints $\mathbf{x}_o = \arg\min_\mathbf{x}/\max f(\mathbf{x})$
- or with constraints

$$\mathbf{x}_o = \arg\min_\mathbf{x}/\max f(\mathbf{x}) \text{ subject to } \mathbf{g}(\mathbf{x}) = 0 \text{ or with } \mathbf{x} \in \mathcal{X}$$

How to solve unconstrained optimization problems?:

1. For certain optimization problems we can find an exact optimum by examining the vanishing point of the derivative i.e. by solving

$$\nabla_\mathbf{x} f(\mathbf{x}) = 0 \text{ where } \nabla_\mathbf{x} f = [\frac{\partial f}{\partial x_1} \dots \frac{\partial f}{\partial x_n}]$$

   $\nabla_\mathbf{x} f$ has the same number of elements as $\mathbf{x}$.

2. Gradient descent. In case that $\nabla_{\mathbf{x}} f(x) = 0$ does not have an analytic solution or the optimum is very computationally expensive to compute by a closed form solution, then gradient descent methods can be applied as alternatives.

   Gradient descent is of the general rule

$$\mathbf{x}^{(t)} = \mathbf{x}^{(t-1)} - \gamma^t \nabla_{\mathbf{x}} f|_{\mathbf{x}=\mathbf{x}^{(t-1)}}$$

   . A very challenging aspect is the computation of the update weight γ^t.

How to solve constraint optimization problems?:

1. Assume the problem

$$\mathbf{x}_o = \arg\min_{\mathbf{x}} f(\mathbf{x})$$

$$\text{s.t. } g(\mathbf{x}) = 0 = \begin{bmatrix} g_1(\mathbf{x}) = 0 \\ \vdots \\ g_k(\mathbf{x}) = 0 \end{bmatrix},$$

   .

   The above optimization problem can be solved by the method of Lagrangian multipliers. In particular, we introduce k Langrangian multipliers (one for each equation $g_j(\mathbf{x}) = 0$, $j = 1, \ldots, k$) and we stack them in a vector $\mathbf{l}$. Then the Lagrangian of the optimization problem can be defined as

$$L(\mathbf{x}, \mathbf{l}) = f(\mathbf{x}) + \mathbf{l}^T g(\mathbf{x})$$

   . We can find optimum $\mathbf{x}_o, \mathbf{l}_o$ by solving $\nabla_{\mathbf{x},\mathbf{l}} L(\mathbf{x}, \mathbf{l}) = 0$ or equivalently

$$\begin{cases} \nabla_{\mathbf{l}} L(\mathbf{x}, \mathbf{l}) = 0 \Rightarrow g(\mathbf{x}) = 0 \\ \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{l}) = 0 \Rightarrow \nabla_{\mathbf{x}} f(\mathbf{x}) = -\nabla_{\mathbf{x}} \mathbf{l}^T g(\mathbf{x}) \end{cases}$$

2. Gradient descent with constraints. In case the method of Lagrangian multipliers cannot be applied, then, a special kind of gradient descent (called project gradient descent) is employed. In particular, the projected gradient descent is given as

$$\mathbf{x}^{(t)} = P_{\mathbf{x}}(\mathbf{x}^{(t-1)} - \gamma^t \nabla_{\mathbf{x}} f|_{\mathbf{x}=\mathbf{x}^{(t-1)}})$$

   and

$$P_{\mathbf{x}} = \arg\min_{\mathbf{y} \in \mathcal{X}} ||\mathbf{x} - \mathbf{y}||^2,$$

   i.e. $P_{\mathbf{x}}$ is projection of $\mathbf{x} \in \mathbb{R}^n$ on to the closest point on $\mathcal{X}$

Finally, we will be using derivatives of functions that employ vectors or matrices as variables such as

- $\nabla_{\mathbf{w}} tr[\mathbf{w}^T \mathbf{A} \mathbf{w}] = 2\mathbf{A}\mathbf{w}$
- $\nabla_{\mathbf{w}} \mathbf{a}^T \mathbf{a} = 2\mathbf{a}$
- $\nabla_{\mathbf{w}} \log |\mathbf{w}| = (\mathbf{w}^{-1})^T$.

No need to remember them by heart (consult to matrix cookbook).