MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering and Computer Science
6.036—Introduction to Machine Learning
Spring Semester 2017

Assignment 0: Preliminaries

Issued: Sunday, February 5
Due: 9am, Friday, February 10

Please submit answers only to questions marked as **(Submit)**

Linear Algebra

1. **Points and Vectors**

   (a) A list of d numbers is a point in a d-dimensional space and also a vector. For 2-dimensional points $[1, 3]$ and $[4, 2]$, produce a diagram showing both interpretations. Use your diagram to show how the vector interpretation facilitates a geometric interpretation of the addition and subtraction of points.

   (b) Give formulae for computing the angle between two vectors, and a vector's length (magnitude). What is the angle between $[0.4, 0.3]$ and $[-0.15, 0.2]$, and what are their respective lengths? Normalize the vectors (so that they both have length 1).

   (c) (**Submit**) Given 3-dimensional vectors $x^{(1)} = [a_1, a_2, a_3]$ and $x^{(2)} = [a_1, -a_2, a_3]$, write down a formula for calculating the angle between them. When is $x^{(1)}$ orthogonal to $x^{(2)}$?

   (d) Explain what it means to project one vector on to another (use a diagram if necessary). What is the projection of $x^{(1)}$ onto $x^{(2)}$, for the $x^{(1)}$ and $x^{(2)}$ given above?

2. **Planes**

   (a) Consider a hyperplane, p_1, in d dimensions. The plane includes all $x = [x_1, \ldots, x_d]$ such that $\theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_d x_d = 0$. We can say that the d-dimensional vector $\theta = [\theta_1, ..., \theta_d]$ together with the offset θ_0 describe the plane p_1. Give another d-dimensional vector θ' and θ'_0 that describe the same plane p_1. How many alternative descriptions are there? Explain.

   (b) Give the equation for determining whether a vector is orthogonal to the plane p_1.

   (c) Explain how to compute the orthogonal projection of a point onto a plane such as p_1.

   (d) (**Submit**) Consider an arbitrary point x, and a hyperplane described by normal $[\theta_1, ..., \theta_d]$ and offset θ_0. The signed distance of x from the plane is the perpendicular distance between x and the plane, multiplied by +1 if x lies on the same side of the plane as the vector θ points and by -1 if x lies on the opposite side. Derive the equation for the signed distance of x from the plane.

   (e) Let p_2 be the plane (a line, since it is 1-dimensional) consisting of the set of points $x = [x_1, x_2]$ for which $3x_1 + x_2 - 1 = 0$.

      i. What is the signed perpendicular distance of point $a = [-1, -1]$ from p_2?

      ii. What is the signed perpendicular distance of the origin from p_2?

      iii. What is the orthogonal projection of point $a = [-1, -1]$ onto p_2?

   (f) Consider a hyperplane in a d-dimensional space. If we project a point onto the plane, can we recover the original point from this projection? If so, show the equation for performing the back-projection. If not, write down an expression for the set of points x that all project to a single point on the plane.

3. **Matrices**

   (a) Consider the matrix:

   $$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 1 & 2 & 1 \end{bmatrix}.$$

      i. Write down A^T.

      ii. Calculate det(A) and det(A^T).

   (b) Evaluate A^{-1} using Python (NumPy). Verify that $AA^{-1} = I$.

   (c) Let g be the row vector $[2, 1, 3]$. Calculate gA.

   (d) Assume C is a 3×2 matrix, and b is a 3×1 (column) vector. What are the dimensions of:

      i. CC^T

      ii. C^TC

      iii. C^Tb

      iv. b^TC

   (e) Write down expressions equivalent to $(AB)^{-1}$ and $(AB)^T$, using only inverses and transposes of A and/or B.

   (f) **(Submit)** Given that $A^T(AB-C) = 0$, where 0 is an $m \times 1$ vector of zeros, derive an expression for B. Assume that all relevant matrices needed for this calculation are invertible.

   (g) What is the rank of:

   $$B = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 4 \\ 5 & 6 & 4 \end{bmatrix}$$

Probability

4. **Probability Density Functions**

   (Submit) Given a continuous random variable $X \in \mathbb{R}$ with probability density function $p_X(x)$[1], which of the following statements are true for $p_X(x)$?

   (a) The value of $p_X(x)$ lies in the interval [0,1].

   (b) When $a < b$, $\int_a^b p_X(x)dx \in [0, 1]$ and represents the probability that the value of random variable X falls within the interval $[a, b]$.

   (c) $p_X(x)$ is always non-negative.

   (d) The integral of $p_X(x)$ from $-\infty$ to ∞ is finite but the specific value may vary.

   (e) $\int_{-\infty}^{\infty} p_X(x)dx = 1$ always.

[1] $p_X(x)$ denotes the probability density function of random variable X evaluated at the value x.

5. **Univariate Gaussians**

   (a) Let $X \sim \mathcal{N}(1, 2)$, i.e., X is a normally distributed random variable with mean 1 and variance 2. What is the probability that $X \in [0.5, 2]$?

   (b) **(Submit)** Let $\mathcal{N}(x; \mu, \sigma^2)$ denote the probability density function for a normally distributed random variable X with mean μ and variance σ^2. Given fixed values of μ and σ, what value of x maximizes $\mathcal{N}(x; \mu, \sigma^2)$? What is the corresponding value of $\mathcal{N}(x; \mu, \sigma^2)$?

   (c) **(Submit)** Suppose a set of points $D = \{x_1, ..., x_n\}$ are drawn independently from some given univariate Gaussian $\mathcal{N}(x; \mu, \sigma^2)$. Provide an expression for the multivariate (joint) probability density function for $x_1, ..., x_n$.

6. **Optimization, gradients**

   Gradient ascent/descent methods are typical tools for maximizing/minimizing functions. Let $L(x, \theta)$ be a function of two vector arguments, $x = [x_1, x_2]^T$ and $\theta = [\theta_1, \theta_2]^T$. We would like to find a value of θ, i.e., both θ_1 and θ_2, such that given x, $L(x, \theta)$ takes its maximum/minimum value. There could be more than one such θ.

   (a) **(Submit)** The gradient $\nabla_\theta L(x, \theta)$ is a vector with two components corresponding to partial derivatives
   $$\frac{\partial}{\partial \theta_j} L(x, \theta), \quad j = 1, 2$$
   Evaluate the gradient when $L(x, \theta) = \log(1 + \exp(-\theta \cdot x))$ where $\theta \cdot x$ is the "dot product" $\theta \cdot x = \theta^T x = \theta_1 x_1 + \theta_2 x_2$.

   (b) **(Submit)** Into which direction does the gradient (viewed as a vector) point? Is the value of $L(x, \theta)$ larger or smaller if we evaluate it at $\theta' = \theta + \epsilon \cdot \nabla_\theta L(x, \theta)$ where ϵ is a small real number?