Statistical Machine Learning (W4400)
Spring 2016
https://courseworks.columbia.edu

John P. Cunningham
jpc2181
Ben Reddy, Phyllis Wan, Ashutosh Nanda
bmr2136, pw2348, an2655

Homework 0

Due: Tuesday 02 February 2016

Students are encouraged to work together, but homework write-ups must be done individually and must be entirely the author's own work. Homework is due at the **beginning** of the class for which it is due. **Late homework will not be accepted under any circumstances.** To receive full credit, students must thoroughly explain how they arrived at their solutions and include the following information on their homeworks: name, UNI, homework number (e.g., HW03), and class (STAT W4400). All homework must be turned in online through Courseworks in PDF format, have a .pdf extension (not zip or other archive!), and be less than 4MB. If programming is part of the assignment, the code must be turned in in one or more .R files. Homeworks not adhering to these requirements will receive no credit. For your convenience (not required), a tex template for producing nice PDF files can be found on courseworks.

Preamble

Prerequisites for this course include a previous course in statistics, elementary probability, multivariate calculus, linear algebra. This homework is designed to allow you to test your background and your ability to adhere to the above submission instructions. All questions have been designed to be solved without any numerical aid – if you are using a computer (which you are welcome to do), you may be missing the didactic purpose.

Submission (50 points)

Submit your homework according to the above instructions. Attention to detail is an essential part of machine learning and the logistics of this course, and penalties for not adhering to the submission instructions will be severe on all homeworks.

Questions (50 points)

In all of the below questions, let:

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \qquad \text{and} \qquad x = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

1. (2 points) What is $B_{2,1}$?

   Solution: 3

2. (2 points) What is $A + B$?

Solution: $\begin{bmatrix} 2 & 4 \\ 5 & 8 \end{bmatrix}$

3. (2 points) What is AB ?

Solution: $\begin{bmatrix} 7 & 10 \\ 14 & 20 \end{bmatrix}$

4. (2 points) What is $rank(A)$?

Solution: 1 (note that $A = [1\ 2]^\top [1\ 2]$)

5. (2 points) What is the largest eigenvalue of A?

Solution: 5 (note that $Az = 5z$ for $z = \alpha[1\ 2]^\top$ for any scalar α)

6. (2 points) What is the eigenvector associated with that largest eigenvalue of A?

Solution: $\frac{1}{\sqrt{5}}[1\ 2]^\top$ (note unit norm)

7. (2 points) What is $|B|$ (determinant of B)?

Solution: -2

8. (2 points) What is $x^\top Ax$?

Solution: 16

9. (2 points) What is $x^\top x$?

Solution: 5

10. (2 points) What is $xx^\top$?

Solution: $\begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}$

11. (2 points) What is $||x||_2$ (the Euclidean norm of x)?

Solution: $\sqrt{5}$

12. (2 points) What is the gradient of $f(x) = x^\top Ax$ w.r.t. x, namely $\nabla_x f(x)$?

Solution: $\begin{bmatrix} 8 \\ 16 \end{bmatrix}$ (note that $\nabla_x f(x) = 2Ax$)

13. (2 points) What is the Hessian of $f(x) = x^\top Ax$ w.r.t. x, namely $\nabla_x^2 f(x)$?

Solution: $\begin{bmatrix} 2 & 4 \\ 4 & 8 \end{bmatrix}$ (note that $\nabla_x^2 f(x) = 2A$)

14. (2 points) We say $x \in I\!R^n$. What is n?

Solution: 2

15. (2 points) I write $y \sim \mathcal{N}(\mu, \sigma^2)$ to denote a Gaussian random variable with mean μ and standard deviation σ. What is $E(y^2)$?

Solution: $E(y^2) = Var(y) + (E(y))^2 = \sigma^2 + \mu^2$

16. (2 points) Say $y \sim \mathcal{N}(2.7, 8)$ and $w \sim \mathcal{N}(3.1, 15)$ are independent random variables. What is the distribution of $y + w$?

Solution: $y + w \sim \mathcal{N}(5.8, 23)$

17. (2 points) I write $\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} -3 \\ 2 \end{bmatrix}, \begin{bmatrix} 4 & \sqrt{3} \\ \sqrt{3} & 3 \end{bmatrix}\right)$ to denote a multivariate Gaussian random variable with dimension $n = 2$, and mean vector μ and covariance matrix Σ as specified. What is the normalizing constant of this distribution?

Solution: $(2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}} = \frac{1}{6\pi}$

18. (2 points) I write $z \sim Bern(p)$ to denote a Bernoulli random variable with bias p. What is the support of z?

Solution: $\{0, 1\}$

19. (2 points) I draw n times independently from $z \sim Bern(p)$. What is the distribution of the number of heads/successes k?

Solution: $P(k \text{ heads}) = \binom{n}{k} p^k (1-p)^{n-k}$

20. (2 points) Find x_1 that maximizes $h(x_1) = \frac{1}{3}x_1^3 - \frac{1}{2}x_1^2 - 6x_1 + \frac{27}{2}$ subject to $x_1 \in [-4, 4]$.

Solution: -2. The critical points are the points where $\frac{\partial h}{\partial x_1} = 0$. Thus $\frac{\partial h}{\partial x_1} = x_1^2 - x_1 - 6$. Using the quadratic formula, the roots are $x = 3$ and $x = -2$. Evaluating these points, $h(3) = 0$, $h(-2) = \frac{125}{6}$. (note the steps thus far are inadequate, as they do not consider the boundary condition). Now $\frac{\partial h}{\partial x_1}$ is convex (opening up), so $h(x_1)$ is decreasing to the left of $x = -2$ and increasing to the right of $x = 3$. Hence the boundary $x = -4$ is not a possible maximum, which we can verify and see that $h(-4) = \frac{49}{6} < h(-2)$. The boundary $x = 4$ has the largest value to the right of $x = 3$, but we can evaluate it to see that $h(4) = \frac{17}{6} < h(-2)$. Thus, $x = -2$ achieves the maximum. (note: the thoroughness of the solution to this simple problem is intentional to demonstrate the level of consideration that should go into your solutions).

21. (2 points) Find the minimum value of $h(x_1)$ subject to the constraint that $x_1 \in [-4, 4]$.

Solution: 0. By the previous solution, the four points to consider are the two critical points and the two boundary points. Of these, $x = 3$ achieves the minimum at $h(3) = 0$.

22. (2 points) Let $\tilde{h}(x_1) = \frac{1}{Z}h(x_1)$; find Z such that $\tilde{h}(x_1)$ has $\int_0^1 \tilde{h}(x_1)dx_1 = 1$.

Solution: $\frac{125}{12}$. Note that $\int_0^1 h(x_1)dx_1 = [\frac{1}{12}x_1^4 - \frac{1}{6}x_1^3 - 3x_1^2 + \frac{27}{2}x_1]|_0^1$. This basic integral is an example of normalization, a critical operation throughout probability and machine learning.

23. (2 points) Let $b(x) = x_1 x_2^3$. Find $\int_{\mathcal{A}} b(x)dx$ for $\mathcal{A} = [0, 3] \times [0, 2]$.

Solution: 18. Note that $\int_{\mathcal{A}} b(x)dx = [[\frac{1}{8}x_1^2x_2^4]_{x_1=0}^{3}]_{x_2=0}^{2}$.

24. (2 points) Let $c(x) = x_1 + \sqrt{3}x_2$; find x that maximizes $c(x)$ subject to $x_1^2 + x_2^2 = 1$.

Solution: $x = \frac{1}{2}\begin{bmatrix}1\\ \sqrt{3}\end{bmatrix}$. The Lagrangian is $\mathcal{L}(x,\lambda) = x_1 + \sqrt{3}x_2 + \lambda(x_1^2 + x_2^2 - 1)$. Setting to 0 the partial derivatives w.r.t. x_1, x_2, and λ, we have: (i) $1 + 2\lambda x_1 = 0$, (ii) $\sqrt{3} + 2\lambda x_2 = 0$, and (iii) $x_1^2 + x_2^2 = 1$. Eliminating λ from (i) and(ii), we see $x_2 = \sqrt{3}x_1$. Substituting that into (iii), we see $x_1^2 + 3x_1^2 = 1 \Rightarrow x_1 = \frac{1}{2} \Rightarrow x_2 = \frac{\sqrt{3}}{2}$.

25. (2 points) Let $g(x) = -x_1 \log x_1 - x_2 \log x_2$; find x that maximizes $g(x)$ subject to $x_1 + x_2 = 1$.

Solution: $x = \frac{1}{2}\begin{bmatrix}1\\ 1\end{bmatrix}$. Setting to 0 the partial derivatives of the Lagrangian $\mathcal{L}(x,\lambda) = g(x) + \lambda(x_1 + x_2 - 1)$, we have: (i) $-\log x_1 - 1 + \lambda = 0$, (ii) $-\log x_2 - 1 + \lambda = 0$, and (iii) $x_1 + x_2 = 1$. From (i) and (ii) we see that $x_1 = x_2$, and thus $x_1 = x_2 = \frac{1}{2}$. Note that this is a simple case of the hugely important fact that a uniform distribution maximizes the entropy of a discrete distribution (you need not know what that sentence means, yet...).

Closing Remark

Linear algebra and optimization are huge and beautiful mathematical fields, and we will only skim the very surface. That said, matrices, vectors, and common manipulations of these objects are the tools of the trade in data science, and thus a basic facility is crucial. Regardless of your ease with the above questions, for linear algebra I recommend studying Zico Kolter's excellent and brief review (for a machine learning class): `http://cs229.stanford.edu/section/cs229-linalg.pdf`. For basic use of Lagrange multipliers, I recommend both of the following:

`www.cs.iastate.edu/%7Ecs577/handouts/lagrange-multiplier.pdf?`.

`www.math.harvard.edu/archive/21a%5Fspring%5F09/PDF/11-08-Lagrange-Multipliers.pdf`

If you feel drastically behind on all these subjects, I strongly recommend serious self-study, including something like the first 16 lectures of:

`http://ocw.mit.edu/courses/mathematics/18-02-multivariable-calculus-fall-2007/index.htm` .

Or perhaps the entirety of: `http://projects.iq.harvard.edu/stat110` .

Successful completion of this course without much of the above background will be challenging, though not impossible.

As a grading rubric, if you have no trouble answering 60-100% of these questions, you are in the right class. If you are able to correctly answer 40-60% of the questions, you will be successful as long as you continue to work to refresh these concepts. If you score well below 40%, you will struggle in this course; you may want to reconsider taking this course without developing more background. These concepts

will be reviewed in brief meaningful detail, but they are the necessary toolkit to begin this material, and thus familiarity is expected.