NYU Center for Data Science: DS-GA 1003

Machine Learning and Computational Statistics (Spring 2018)

Brett Bernstein[*], David Rosenberg, Ben Jakubowski

January 23, 2018

Instructions: Following most lab and lecture sections, we will be providing concept checks for review. Each concept check will:

- List the lab/lecture learning objectives. You will be responsible for mastering these objectives, and demonstrating mastery through homework assignments, exams (midterm and final), and on the final course project.
- Include concept check questions. These questions are intended to reinforce the lab/lectures, and help you master the learning objectives.

You are strongly encourage to complete all concept check questions, and to discuss these (and related) problems on Piazza and at office hours. However, problems marked with a ($\star$) are considered optional.

Pre-Lecture 2: Optimization and linear algebra

Instructions: Prior to lecture 2, please review the following problems

Optimization Prerequisites for Lasso

1. Given $a \in \mathbb{R}$ we define a^+, a^- as follows:

$$a^+ = \begin{cases} a & \text{if } a \geq 0, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad a^- = \begin{cases} -a & \text{if } a < 0, \\ 0 & \text{otherwise.} \end{cases}$$

   We call a^+ the *positive part* of a and a^- the *negative part* of a. Note that $a^+, a^- \geq 0$.

   (a) Give an expression for a in terms of a^+, a^-.

   (b) Give an expression for $|a|$ in terms of a^+, a^-.

[*]Brett authored these concept checks for Spring 2017 DS-GA 1003, and the work is almost entirely his. Later (minor) modifications were made by David Rosenberg and Ben Jakubowski.

For $x \in \mathbb{R}^d$ define $x^+ = (x_1^+, \ldots, x_d^+)$ and $x^- = (x_1^-, \ldots, x_d^-)$.

(c) Give an expression for x in terms of x^+, x^-.

(d) Give an expression for $\|x\|_1$ without using any summations or absolute values. [Hint: Use x^+, x^- and the vector $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{R}^d$.]

Solution.

(a) $a = a^+ - a^-$

(b) $|a| = a^+ + a^-$

(c) $x = x^+ - x^-$

(d) $\|x\|_1 = \mathbf{1}^T(x^+ + x^-)$

2. Let $f : \mathbb{R} \to \mathbb{R}$ and $S \subseteq \mathbb{R}$. Consider the two optimization problems

$$\begin{array}{ll} \text{minimize}_{x\in\mathbb{R}} & |x| \\ \text{subject to} & f(x) \in S \end{array} \qquad \text{and} \qquad \begin{array}{ll} \text{minimize}_{a,b\in\mathbb{R}} & a+b \\ \text{subject to} & f(a-b) \in S \\ & a, b \geq 0. \end{array}$$

Solve the following questions.

(a) If x in the first problem satistfies $f(x) \in S$ show how to quickly compute (a, b) for the second problem with $a + b = |x|$ and $f(a - b) \in S$.

(b) If a, b in the second problem satisfy $f(a - b) \in S$, show how to quickly compute an x for the first problem with $|x| \leq a + b$ and $f(x) \in S$.

(c) Assume x is a minimizer for the first problem with minimum value p_1^* and (a, b) is a minimizer for the second problem with minimum p_2^*. Using the previous two parts, conclude that $p_1^* = p_2^*$.

Solution.

(a) Let $a = x^+$ and $b = x^-$. Then $a + b = |x|$ and $a - b = x$.

(b) Let $x = a - b$ and note that $|x| = |a - b| \leq |a| + |b| = a + b$.

(c) Part a) shows $p_2^* \leq p_1^*$ by letting $\hat{a} = x^+$ and $\hat{b} = x^-$. Part b) shows $p_1^* \leq p_2^*$ by letting $\hat{x} = a - b$.

3. Let $f : \mathbb{R}^d \to \mathbb{R}$, $S \subseteq \mathbb{R}$ and consider the following optimization problem:

$$\begin{array}{ll} \text{minimize}_{x\in\mathbb{R}^d} & \|x\|_1 \\ \text{subject to} & f(x) \in S, \end{array}$$

where $\|x\|_1 = \sum_{i=1}^d |x_i|$. Give a new optimization problem with a linear objective function and the same minimum value. Show how to convert a solution to your new problem into a solution to the given problem. [Hint: Use the previous two problems.]

Solution. Consider the minimization problem

$$\begin{array}{ll} \text{minimize}_{a,b\in\mathbb{R}^d} & \mathbf{1}^T(a+b) \\ \text{subject to} & f(a-b)\in S, \\ & a_i, b_i \geq 0 \quad \text{for } i=1,\ldots,d. \end{array}$$

Let p_1^* be the minimum for the original problem, and p_2^* the minimum for our new problem. We first show $p_1^* = p_2^*$. Suppose x is a minimizer for the original problem and let $a = x^+$ and $b = x^-$. Then by the first question $\mathbf{1}^T(a+b) = \|x\|_1$ and $a-b = x$. This shows $p_2^* \leq p_1^*$. Next suppose (a,b) is a minimizer for our new problem, and let $x = a - b$. Then

$$\|x\|_1 = \|a-b\|_1 = \sum_{i=1}^{d} |a_i - b_i| \leq \sum_{i=1}^{d} |a_i| + |b_i| = \sum_{i=1}^{d} a_i + b_i = \mathbf{1}^T(a+b).$$

This proves $p_1^* \leq p_2^*$.

Finally, given a minimizer (a,b) for the new problem we recover a minimizer x for the original problem by letting $x = a - b$.

Linear Algebra Prerequisites for Linear Regressions

1. When performing linear regression we obtain the *normal equations* $A^TAx = A^Ty$ where $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, and $y \in \mathbb{R}^m$.

   (a) If $\mathbf{rank}(A) = n$ then solve the normal equations for x.

   (b) ($\star$) What if $\mathbf{rank}(A) \neq n$?

   Solution.

   (a) We first show that $\mathbf{rank}(A^TA) = n$ to show that we can invert A^TA. By the rank-nullity theorem, we can do this by showing A^TA has trivial nullspace. Note that for any $x \in \mathbb{R}^n$ we have

   $$A^TAx = 0 \implies x^TA^TAx = 0 \implies \|Ax\|_2^2 = 0 \implies Ax = 0 \implies x = 0.$$

   This last implication follows since $\mathbf{rank}(A) = n$ so A has trivial nullspace (again by rank-nullity). This proves A^TA has a trivial nullspace, and thus A^TA is invertible. Applying the inverse we obtain

   $$x = (A^TA)^{-1}A^Ty.$$

   Since A^TA is invertible, our answer for x is unique.

(b) We will show that the equation always has infinitely many solutions x. First note that $\mathbf{rank}(A) \neq n$ implies $\mathbf{rank}(A) < n$ since you cannot have larger rank than the number of columns. Next, recall $\mathbf{rank}(A) = \mathbf{rank}(A^TA)$. Hence, by rank-nullity, A^TA has a non-trivial nullspace, which in turn implies that if there is a solution, there must be infinitely many solutions.

Next note A^T and A^TA have the same column space. To see this, first note that every vector of the form A^TAx must be a linear combination of the columns of A^T, and thus lies in the column space of A^T. Since $\mathbf{rank}(A^TA) = \mathbf{rank}(A) = \mathbf{rank}(A^T)$, this implies A^T and A^TA have the same column spaces.
A specific solution can be computed as $x = (A^TA)^+A^Ty$, where $(A^TA)^+$ is the *pseudoinverse* of A^TA. Of the infinitely many possible solutions x, this gives the one that minimizes $\|x\|_2$. More precisely, $x = (A^TA)^+A^Ty$ solves the optimization problem

$$\begin{array}{ll} \text{minimize} & \|x\|_2 \\ \text{subject to} & A^TAx = A^Ty. \end{array}$$

2. Prove that $A^TA + \lambda\mathbf{I}_{n\times n}$ is invertible if $\lambda > 0$ and $A \in \mathbb{R}^{n\times n}$.

Solution. If $(A^TA + \lambda\mathbf{I}_{n\times n})x = 0$ then

$$0 = x^T(A^TA + \lambda\mathbf{I}_{n\times n})x = \|Ax\|_2^2 + \lambda\|x\|_2^2 \implies x = 0.$$

Thus $A^TA + \lambda\mathbf{I}_{n\times n}$ has trivial nullspace. Alternatively, we could notice that A^TA is positive semidefinite, so adding $\lambda\mathbf{I}_{n\times n}$ will give a matrix whose eigenvalues are all at least $\lambda > 0$. A square matrix is invertible iff its eigenvalues are all non-zero.

3. ($\star$) Describe the following set geometrically:

$$\left\{ v \in \mathbb{R}^2 \;\middle|\; v^T \begin{pmatrix} 2 & 2 \\ 0 & 2 \end{pmatrix} v = 4 \right\}.$$

Solution. The set is an ellipse with semi-axis lengths $2/\sqrt{3}$ and 2 rotated counter-clockwise by $\pi/4$. Letting $v = (x, y)^T$ and multiplying all terms we get

$$2x^2 + 2xy + 2y^2 = 4.$$

From precalculus we can see this is a conic section, and must be an ellipse or a hyperbola, but more work is needed to determine which one. Instead of proceeding along these lines, let's use linear algebra to give a cleaner treatment that extends to higher dimensions.
Let $A = \begin{pmatrix} 2 & 2 \\ 0 & 2 \end{pmatrix}$. Since v^TAv is a number, we must have $(v^TAv)^T = v^TA^Tv$. This gives

$$v^TA^Tv = v^TAv = \frac{1}{2}v^T(A^T + A)v = v^T \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} v.$$

Our new matrix is symmetric, and thus allows us to apply the spectral theorem to diagonalize it with an orthonormal basis of eigenvectors. In other words, by rotating our axes we can get a diagonal matrix. Either doing this by hand, or using a computer (Matlab, Mathematica, Numpy) we obtain

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = Q \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} Q^T \quad \text{where} \quad Q = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix}.$$

The set

$$\left\{ w \in \mathbb{R}^2 \;\middle|\; w^T \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} w = 4 \right\}$$

is an ellipse with semi-axis lengths $2/\sqrt{3}$ and 2 since it corresponds to the equation $3w_1^2 + w_2^2 = 4$. Since Q performs a counter-clockwise rotation by $\pi/4$ we obtain the answer. More concretely,

$$w^T \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} w = 4 \iff (Qw)^T Q \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} Q^T (Qw) = 4 \iff (Qw)^T \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} (Qw) = 4$$

so

$$\{v \mid v^T A v = 4\} = \left\{ Qw \;\middle|\; w^T \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} w = 4 \right\}.$$

Figure 1: Rotated Ellipse

More generally, the solution to $v^T A v = c$ for $v \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$ and $c > 0$ will be an ellipsoid if A is positive definite. The ith semi-axis will have length $\sqrt{c/\lambda_i}$ where λ_i is the ith eigenvalue of A.