Machine Learning for Finance – Problem Set 1

Neal Parikh

January 30, 2018

Instructions. Do not refer to any outside sources to complete this assignment, in accordance with the honor code. If you discussed any problems with other students, indicate that in your solutions.

1. *Matrix algebra.* Let $A \in \mathbf{R}^{m \times p}$ and let $a_i \in \mathbf{R}^m$ and $\tilde{a}_j^T \in \mathbf{R}^p$ refer to the columns and rows of A, respectively. Let $B \in \mathbf{R}^{p \times n}$ and define b_i and $\tilde{b}_j^T$ as its columns and rows, respectively. In general, vectors $x \in \mathbf{R}^n$ refer to *column* vectors, and column vectors can be specified entrywise in either of the following two ways:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad x = (x_1, x_2, \ldots, x_n).$$

   In other words, the notation $(x_1, \ldots, x_n)$ specifies a column vector in $\mathbf{R}^n$.

   (a) Express the matrix-vector product Ax in terms of the columns of A.

   (b) Express the matrix-vector product Ax in terms of the rows of A.

   (c) Express the matrix product AB in terms of the rows of A and the columns of B.

   (d) Express the matrix product AB in terms of the columns of A and the rows of B.

   (e) *Least squares.* Let $A \in \mathbf{R}^{m \times n}$ and $b \in \mathbf{R}^m$. We will see that

$$\hat{x} = \operatorname*{argmin}_x \|Ax - b\|_2^2$$

   can be expressed in closed form as

$$\hat{x} = (A^T A)^{-1} A^T b$$

   when the columns of A are linearly independent. Express $\hat{x}$ in terms of the rows of A.

2. *Representing linear functions.* For each description of y below, express it as $y = Ax$ for some A. (You should specify A.)

   (a) y_i is the difference between x_i and the average of $x_1, \ldots, x_{i-1}$. (We take $y_1 = x_1$.)

(b) y_i is the difference between x_i and the average value of all other x_j's, *i.e.*, the average of $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$.

3. *Rank and null space.* Let $A \in \mathbf{R}^{m \times n}$. Recall that $\mathbf{rank}\, A$ is the (maximal) number of linearly independent rows or columns of A. We say that A has *full rank* if $\mathbf{rank}\, A = \min(m, n)$, and that A is *rank deficient* otherwise.

   The *null space* or *kernel* of A is defined as

   $$\mathcal{N}(A) = \{x \in \mathbf{R}^n \mid Ax = 0\}.$$

   The sum of the rank and dimension of the null space of A equals n, the number of columns.

   Let $z \in \mathbf{R}^n$. What are the rank and null space of zz^T?

4. *Positive definite matrices.* Let $\mathbf{S}^n \subset \mathbf{R}^{n \times n}$ be the set of symmetric $n \times n$ matrices. A matrix $A \in \mathbf{S}^n$ is *positive semidefinite*, denoted $A \succeq 0$ or $A \in \mathbf{S}^n_+$, if $x^T Ax \geq 0$ for all nonzero $x \in \mathbf{R}^n$. If $x^T Ax > 0$ for all nonzero x, then A is *positive definite*, denoted $A \succ 0$ or $A \in \mathbf{S}^n_{++}$. A matrix A is negative (semi)definite if $-A$ is positive (semi)definite.

   (a) Show that the identity matrix I is positive definite.

   (b) Show that zz^T, where $z \in \mathbf{R}^n$, is positive semidefinite but not positive definite.

   (c) Show that if $A \in \mathbf{S}^n$ is either positive or negative definite, then A is full rank.

   (d) Given any matrix $B \in \mathbf{R}^{m \times n}$, show that the *Gram matrix* $G = B^T B$ is positive semidefinite. In addition, show that if $m \geq n$ and B is full rank, then G is positive definite.

5. *Risk models.* Let $x \in \mathbf{R}^n$ represent a portfolio of assets, with x_i representing the amount held of asset i. In portfolio theory, the risk of a portfolio is typically represented as the quadratic form $x^T \Sigma x$, where $\Sigma \succeq 0$ is called the *risk model*.

   The idea is that if Σ_{ij} is large, then we expect assets i and j to go up or down together, while if Σ_{ij} is negative then we expect assets i and j to move in opposite directions. A well-diversified portfolio would not place too much in both assets i and j if Σ_{ij} is positive, since they are expected to behave similarly.

   Give a financial interpretation of Σ not being positive definite.

6. *Derivatives and gradients.* Suppose $f : \mathbf{R}^n \to \mathbf{R}^m$ and $x \in \mathbf{int}\, \mathbf{dom}\, f$. The function f is differentiable at x if there exists a matrix $Df(x) \in \mathbf{R}^{m \times n}$ that satisfies

   $$\lim_{z \in \mathbf{dom}\, f, z \neq x, z \to x} \frac{\|f(z) - f(x) - Df(x)(z - x)\|_2}{\|z - x\|_2} = 0, \tag{1}$$

   in which case $Df(x)$ is referred to as the *derivative* or *Jacobian* of f at x. (There can be at most one matrix that satisfies (1).) The function f is *differentiable* if $\mathbf{dom}\, f$ is open and it is differentiable at every point in its domain.

   Note that $Df(x)$ is a linear map $Df(x) : \mathbf{R}^n \to \mathbf{R}^m$. The affine function of $z \in \mathbf{R}^n$ given by

   $$f(x) + Df(x)(z - x)$$

is called the *first-order approximation* of f at x. This function agrees with f at $z = x$; when z is close to x, this approximation is very close to f.

The derivative is typically found by computing partial derivatives of f:

$$Df(x)_{ij} = \frac{\partial f_i}{\partial x_j}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n.$$

We will mostly encounter the special case when f is real-valued, *i.e.*, $f : \mathbf{R}^n \to \mathbf{R}$. In this case, the derivative $Df(x)$ is a $1 \times n$ matrix, *i.e.*, a row vector. Its transpose is called the *gradient* of the function and is denoted

$$\nabla f(x) = Df(x)^T,$$

which is a column vector in $\mathbf{R}^n$ with the partial derivatives of f as its components. The first-order approximation at $x \in \mathbf{int}\,\mathbf{dom}\, f$ can be expressed as

$$f(x) + \nabla f(x)^T (z - x),$$

an affine function of z.

(a) *Quadratic function.* Compute the gradient of the function $f : \mathbf{R}^n \to \mathbf{R}$ given by

$$f(x) = (1/2)x^T P x + q^T x + r,$$

where $P \in \mathbf{S}^n$, $q \in \mathbf{R}^n$, and $r \in \mathbf{R}$.

(b) *Chain rule.* Let $f : \mathbf{R}^n \to \mathbf{R}^m$, $g : \mathbf{R}^m \to \mathbf{R}^p$, and $h(x) = g(f(x))$. Suppose that f is differentiable at $x \in \mathbf{int}\,\mathbf{dom}\, f$ and g is differentiable at $f(x) \in \mathbf{int}\,\mathbf{dom}\, g$; then h is differentiable at x, with derivative

$$Dh(x) = Dg(f(x))Df(x).$$

Let $f : \mathbf{R}^n \to \mathbf{R}$, $A \in \mathbf{R}^{n \times p}$, $b \in \mathbf{R}^n$, and let $g(x) = f(Ax + b)$. Express $\nabla g(x)$ in terms of A, b, and $\nabla f(x)$.

7. *Hessians.* The second derivative or *Hessian matrix* of $f : \mathbf{R}^n \to \mathbf{R}$ at $x \in \mathbf{int}\,\mathbf{dom}\, f$, denoted $\nabla^2 f(x)$, is given by

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n,$$

where the partial derivatives are evaluated at x (and are assumed to exist).

The Hessian can also be expressed as

$$\nabla^2 f(x) = D\nabla f(x),$$

where $\nabla f : \mathbf{R}^n \to \mathbf{R}^n$ is called the *gradient mapping* of f, defined by $\nabla f : x \mapsto \nabla f(x)$ when f is differentiable.

(a) *Quadratic function.* Compute the Hessian matrix of

$$f(x) = (1/2)x^T Ax + b^T x,$$

where $A \in \mathbf{S}^n$ and $b \in \mathbf{R}^n$.

(b) *Affine composition.* Let $f : \mathbf{R}^n \to \mathbf{R}$, $A \in \mathbf{R}^{n \times p}$, $b \in \mathbf{R}^n$, and let

$$g(x) = f(Ax + b).$$

Find the Hessian of g.

8. *Medical diagnosis.* After your yearly checkup, the doctor has good news and bad news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (*i.e.*, the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, affecting only one in 10,000 people.

   What are the chances that you have the disease? Why is it good that the disease is rare?

9. *Curse of dimensionality.* Consider a dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, with $x_i \in \mathbf{R}^n$ and $y_i \in \mathbf{R}$. Given a new query point x^{new} whose (unknown) label we want to predict, the *k-nearest neighbors* method predicts the average of the labels y_i of the k points x_i in $\mathcal{D}$ closest to x^{new}, where 'closest' is in terms of Euclidean distance if not specified.

   It might seem like if N is sufficiently large, then it would always be possible to come up with good predictions with this method since we can always find a fairly large neighborhood of observations near the query point x^{new} and average them. Unfortunately, this intuition breaks down in high dimensions – *i.e.*, when n is relatively large – which is the case for most modern problems. This phenomenon is known as the *curse of dimensionality*, a phrase coined by the mathematician Richard Bellman in the early 1960s.

   For example, the idea that N will be sufficiently large is itself problematic. Suppose we consider that $N = 100$ is a sufficiently dense sample when $n = 1$, *i.e.*, a single-dimensional problem. Since the sampling density is proportional to $N^{1/n}$, we would need a sample size of $N = 100^{10}$ to achieve the same sampling density for even a ten dimensional problem. In other words, in high dimensions, all realistic training sets sparsely populate the input space.

   This problem will consider some other aspects of the curse of dimensionality, namely that we end up needing to look at extremely large 'neighborhoods' in order to capture a given amount of nearby data, and the notion of 'nearness' behaves strangely because all the points wind up on the boundary of the sample and so appear to be equidistant.

   Note. Below, you can use the fact that the volume of a hypersphere of radius r in n dimensions is given by

$$V(r, n) = \frac{\pi^{n/2}}{(n/2)!} r^n$$

   when n is even. (When n is odd, the factorial is replaced with a gamma function, but this is not relevant for the problem.)

(a) Consider using k-nearest neighbors for points uniformly distributed in an n-dimensional unit hypercube. If we want to capture a fraction ρ of the observations in a hypercubical neighborhood around a given point, what should the edge length of this neighborhood be as a function of n and ρ?

Evaluate this function for $(\rho, n) \in \{(0.01, 10), (0.1, 10), (0.01, 1000)\}$.

(b) Consider a sphere of radius r in $\mathbf{R}^n$. Show that the fraction of the sphere's volume in the surface shell lying at values of the radius between $r - \epsilon$ and r, where $0 < \epsilon < r$, is

$$\rho = 1 - \left(1 - \frac{\epsilon}{r}\right)^n.$$

Evaluate ρ for the cases $n \in \{2, 10, 1000\}$, with $\epsilon/r \in \{0.01, 0.5\}$.

(c) Consider a unit cube in $\mathbf{R}^n$ with an inscribed sphere. What are the volumes of the cube and the sphere as $n \to \infty$? (You can verify this empirically; a proof is not necessary.)

(d) Let $x_1, \ldots, x_N \in \mathbf{R}^n$ be uniformly distributed in an n-dimensional unit sphere centered at the origin. Show that the median distance from the origin to the closest data point is given by the expression

$$d(N, n) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/n}.$$

Evaluate d for $N = 500$, $n = 10$.