Machine Learning – Brett Bernstein

Week 1 Lab: Concept Check Exercises

Multivariable Calculus Exercises

1. If $f'(x;u) < 0$ show that $f(x+hu) < f(x)$ for sufficiently small $h > 0$.

   Solution. The directional derivative is given by

   $$f'(x;u) = \lim_{h\to 0} \frac{f(x+hu)-f(x)}{h} < 0.$$

   By the definition of a limit, there must be a $\delta > 0$ such that

   $$\frac{f(x+hu)-f(x)}{h} < 0$$

   whenever $|h| < \delta$. If we restrict $0 < h < \delta$ then we have

   $$f(x+hu)-f(x) < 0 \implies f(x+hu) < f(x)$$

   as required.

2. Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable, and assume that $\nabla f(x) \neq 0$. Prove

   $$\arg\max_{\|u\|_2=1} f'(x;u) = \frac{\nabla f(x)}{\|\nabla f(x)\|_2} \quad \text{and} \quad \arg\min_{\|u\|_2=1} f'(x;u) = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}.$$

   Solution. By Cauchy-Schwarz we have, for $\|u\|_2 = 1$,

   $$|f'(x;u)| = |\nabla f(x)^T u| \leq \|\nabla f(x)\|_2 \|u\|_2 = \|\nabla f(x)\|_2.$$

   Note that

   $$\nabla f(x)^T \frac{\nabla f(x)}{\|\nabla f(x)\|_2} = \|\nabla f(x)\|_2 \quad \text{and} \quad \nabla f(x)^T \frac{-\nabla f(x)}{\|\nabla f(x)\|_2} = -\|\nabla f(x)\|_2,$$

   so these achieve the maximum and minimum bounds given by Cauchy-Schwarz.

3. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x,y) = x^2 + 4xy + 3y^2$. Compute the gradient $\nabla f(x,y)$.

   Solution. Computing the partial derivatives gives

   $$\partial_1 f(x,y) = 2x + 4y \quad \text{and} \quad \partial_2 f(x,y) = 4x + 6y.$$

   Thus the gradient is given by

   $$\nabla f(x,y) = \begin{pmatrix} 2x+4y \\ 4x+6y \end{pmatrix}.$$

4. Compute the gradient of $f : \mathbb{R}^n \to \mathbb{R}$ where $f(x) = x^T Ax$ and $A \in \mathbb{R}^{n\times n}$ is any matrix.

   Solution. Here we show two methods. In either case we can obtain differentiability by noticing the partial derivatives are continuous.

   (a) Since
   $$f(x) = x^T Ax = \sum_{i,j=1}^{n} a_{ij} x_i x_j$$
   we have
   $$\partial_k f(x) = \sum_{j=1}^{n} (a_{kj} + a_{jk}) x_j$$
   so
   $$\nabla f(x) = (A + A^T)x.$$

   (b) Note that
   $$\begin{aligned} f(x+tv) &= (x+tv)^T A(x+tv) \\ &= x^T Ax + tx^T Av + tv^T Ax + t^2 v^T Av \\ &= f(x) + t(x^T A + x^T A^T)v + t^2(v^T Av). \end{aligned}$$
   Thus
   $$f'(x;v) = \lim_{t\to 0} \frac{f(x+tv) - f(x)}{t} = \lim_{t\to 0} (x^T A + x^T A^T)v + t(v^T Av) = (x^T A + x^T A^T)v.$$
   This shows
   $$\nabla f(x) = (A + A^T)x.$$

5. Compute the gradient of the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ given by
   $$f(x) = b + c^T x + x^T Ax,$$
   where $b \in \mathbb{R}$, $c \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$.

   Solution. First consider the linear function $g(x) = c^T x$. Note that
   $$g(x+tv) = c^T(x+tv) = c^T x + tc^T v \implies \nabla f(x) = c.$$
   As the derivative is linear we can combine this with the previous problem to obtain
   $$\nabla f(x) = c + (A + A^T)x.$$

6. Fix $s \in \mathbb{R}^n$ and consider $f(x) = (x-s)^T A(x-s)$ where $A \in \mathbb{R}^{n\times n}$. Compute the gradient of f.

   Solution. We give two methods.

(a) Let $g(x) = x^T Ax$ and $h(x) = x - s$ so that $f(x) = g(h(x))$. By the vector-valued form of the chain rule we have

$$\nabla f(x) = \nabla g(h(x))^T Dh(x) = (A + A^T)(x - s),$$

where $Dh(x) = \mathbf{I}_{n\times n}$ is the Jacobian matrix of h.

(b) We have

$$(x - s)^T A(x - s) = x^T Ax - s^T(A + A^T)x + s^T As.$$

Computing the gradient gives

$$\nabla f(x) = (A + A^T)x - (A + A^T)s = (A + A^T)(x - s).$$

7. Consider the ridge regression objective function

$$f(w) = \|Aw - y\|_2^2 + \lambda\|w\|_2^2,$$

where $w \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $y \in \mathbb{R}^m$, and $\lambda \in \mathbb{R}_{\geq 0}$.

(a) Compute the gradient of f.

(b) Express f in the form $f(w) = \|Bw - z\|_2^2$ for some choice of B, z.

(c) Using either of the parts above, compute

$$\arg\min_{w\in\mathbb{R}^n} f(w).$$

Solution.

(a) We can express $f(w)$ as

$$f(w) = (Aw - y)^T(Aw - y) + \lambda w^T w = w^T A^T Aw - 2y^T Aw + y^T y + \lambda w^T w.$$

Applying our previous results gives (noting $w^T w = w^T \mathbf{I}_{n\times n} w$)

$$\nabla f(w) = 2A^T Aw - 2A^T y + 2\lambda w = 2(A^T A + \lambda \mathbf{I}_{n\times n})w - 2A^T y.$$

(b) Let

$$B = \begin{pmatrix} A \\ \sqrt{\lambda}\mathbf{I}_{n\times n} \end{pmatrix} \quad \text{and} \quad z = \begin{pmatrix} y \\ \mathbf{0}_{n\times 1} \end{pmatrix}$$

written in block-matrix form.

(c) The argmin is $w = (A^T A + \lambda \mathbf{I}_{n\times n})^{-1} A^T y$. To see why the inverse is valid, see the linear algebra questions below.

8. Compute the gradient of

$$f(\theta) = \lambda\|\theta\|_2^2 + \sum_{i=1}^{n} \log(1 + \exp(-y_i \theta^T x_i)),$$

where $y_i \in \mathbb{R}$ and $\theta \in \mathbb{R}^m$ and $x_i \in \mathbb{R}^m$ for $i = 1, \ldots, n$.

Solution. As the derivative is linear, we can compute the gradient of each term separately and obtain

$$\nabla f(\theta) = 2\lambda\theta - \sum_{i=1}^{n} \frac{\exp(-y_i\theta^T x_i)}{1+\exp(-y_i\theta^T x_i)} y_i x_i,$$

where we used the techniques from Recitation 1 to differentiate the log terms.

Linear Algebra Exercises

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) 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. 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. We will show that A^T and A^TA have the same column space. This will imply A^Ty is in the column space of A^TA giving the result. 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. Above we proved that the column space of A^TA has dimension n, the same as the column space of A^T (since $\mathbf{rank}(A^T) = \mathbf{rank}(A)$). Thus 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. 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 $(v^TAv)^T = v^TAv$. 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

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

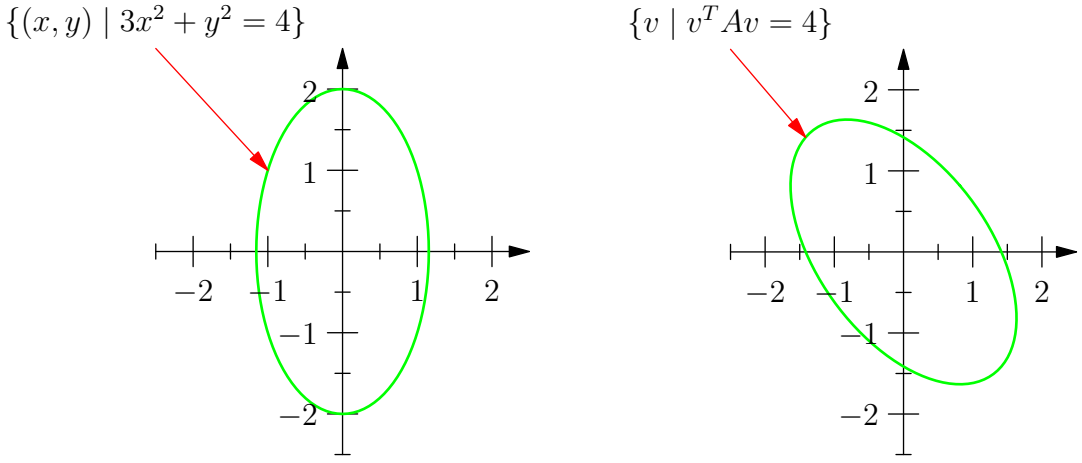

Figure 1: Rotated Ellipse

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.