6740: Computational Data Analysis

How to use (and develop judgement for) machine learning methods

Discussion 1

Jan 20

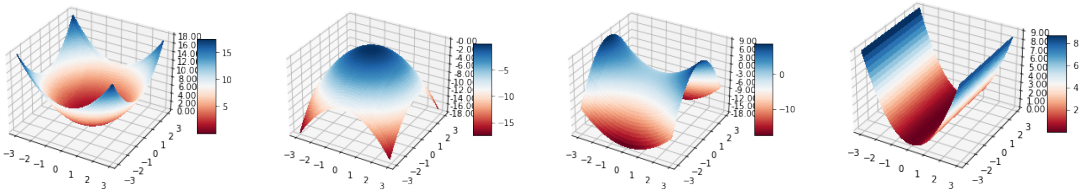
Objective. Calculus, linear algebra and probability review. This should also give you a flavor for what is to come next week, and prepare you for portions of HW1.

Problem 1 (Calculus and linear algebra): Quadratic functions of vectors.

(a) **Consider the vector-valued function $f(x) = x^\top Ax + b^T x + c$ for a square matrix A and scalar c. Geometrically, what do you expect f to look like (say when $b = 0$)? What are the various cases here? Think first about the one-dimensional case, and then about the two-dimensional case, and then generalize.**

Let's think about the one-dimensional case, where x, A, b, c are all scalars. This is just a quadratic function of a single variable, and is a bowl-shaped function with some minimum if $A > 0$ and an inverted bowl if $A < 0$. The location of the minimum is determined by b and c, but the curvature only by A: the larger A is in magnitude, the more curved the function is. If $A = 0$, then this is just a linear function with slope b and intercept c.

Let's try and generalize this intuition to two-dimensions, first by considering the case where A is a diagonal matrix. In other words, we can write $f(x) = (A_{11}x_1^2 + b_1x_1 + c_1) + (A_{22}x_2^2 + b_2x_2 + c_2)$, i.e., the function is just a sum of two single-dimensional quadratic functions, and we have multiple cases (8, to be precise) depending on the individual values of (A_{11}, A_{22}). Let's cover a few of these: the function is bowl-shaped if both values are positive, it is an inverted bowl if both values are negative, and it is some combination of the two if one value is positive and the other negative. This is illustrated in the figure below. Notice closely what happens if one of the two values is zero: we simply have a "flat" quadratic function.



$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \qquad A = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix} \qquad A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Figure 1. Examples of four quadratic functions in two variables (all have $b = 0$, $c = 0$. Fill in the axis labels to convince yourself you understand.

Now what if A is no longer diagonal? Well the first step is to notice that we can assume

without loss of generality that A is symmetric: let $Q = \frac{1}{2}(A + A^\top)$ and note that

$$Q^\top = \frac{1}{2}(A^\top + A) = Q$$

so Q is symmetric, and

$$x^\top Q x = x^\top \frac{A + A^\top}{2} x = \frac{1}{2} x^\top A x + \frac{1}{2} x^\top A^\top x = x^\top A x$$

so $f(x) = x^\top Q x + b^\top x + c$. Ok, so it suffices to consider symmetric matrices A. But how does this help us?

Notice that every symmetric matrix A has an *eigenvalue decomposition*, i.e., it can be written as $A = U \Lambda U^\top$ for a diagonal matrix Λ of eigenvalues and an *orthonormal* matrix U. (Note that $U^\top$ is also an orthonormal matrix, and that $U^{-1} = U^\top$ (why?). What is the operational interpretation of U? It always *rotates* (or reflects) vectors, performing a *change of coordinates*! (Think of what the matrix $U = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ does) In other words, letting $y = U^\top x$, we can now write $f(x) = g(y) = y^\top \Lambda y + b^\top U y + c$, and so we now have a different function in which the quadratic term is once again given by a diagonal matrix. In fact, if $b = 0$, the value of the function is exactly the same (think about why). So our graphs from before also capture this case. You have been provided a Jupyter notebook with code to generate these quadratic plots. Play with it and see if you understand; feel free to add nonzero values of b and c.

In the case where we operate in higher dimensions, this intuition extends exactly as is: there are "positive" directions of curvature and negative directions of curvature, and the function could be flat along some directions.

(b) **When does the function f above have a finite minimum? When is this minimum unique?**

As explained above, the minimum is finite if all eigenvalues of A are non-negative (in other words, if the matrix is positive semidefinite), otherwise, there is some direction of negative curvature along which we can escape to $-\infty$.

(c) **What is the gradient of f with respect to x?**

Similarly to Question (a), we can assume that A is symmetric, i.e., $A_{i,j} = A_{j,i}$. We have

$$\begin{aligned} f(x) &= x^\top A x + b^\top x + c \\ &= x^\top \begin{pmatrix} \sum_{j=1}^n A_{1,j} x_j \\ \vdots \\ \sum_{j=1}^n A_{n,j} x_j \end{pmatrix} + \sum_{i=1}^n b_i x_i + c \\ &= \sum_{i=1}^n \sum_{j=1}^n A_{i,j} x_i x_j + \sum_{i=1}^n b_i x_i + c. \end{aligned}$$

We want to compute the gradient $\nabla f(x)$, i.e., to compute each partial derivative $\frac{\partial f}{\partial x_k}(x)$, since

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}.$$

We have

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}x_i x_j &= \sum_{\substack{i=1\\ i\neq k}}^{n}\sum_{j=1}^{n} A_{i,j}x_i x_j + \sum_{j=1}^{n} A_{k,j}x_k x_j \\
&= \sum_{\substack{i=1\\ i\neq k}}^{n}\sum_{\substack{j=1\\ j\neq k}}^{n} A_{i,j}x_i x_j + \sum_{\substack{i=1\\ i\neq k}}^{n} A_{i,k}x_i x_k + \sum_{\substack{j=1\\ j\neq k}}^{n} A_{k,j}x_k x_j + A_{k,k}x_k^2 \\
&= \sum_{\substack{i=1\\ i\neq k}}^{n}\sum_{\substack{j=1\\ j\neq k}}^{n} A_{i,j}x_i x_j + \sum_{\substack{i=1\\ i\neq k}}^{n} A_{k,i}x_k x_i + \sum_{\substack{j=1\\ j\neq k}}^{n} A_{k,j}x_k x_j + A_{k,k}x_k^2 \\
&= \sum_{\substack{i=1\\ i\neq k}}^{n}\sum_{\substack{j=1\\ j\neq k}}^{n} A_{i,j}x_i x_j + 2\sum_{\substack{j=1\\ j\neq k}}^{n} A_{k,j}x_k x_j + A_{k,k}x_k^2.
\end{aligned}$$

Thus,

$$f(x) = \underbrace{\sum_{\substack{i=1\\ i\neq k}}^{n}\sum_{\substack{j=1\\ j\neq k}}^{n} A_{i,j}x_i x_j}_{\text{independent of } x_k} + 2\underbrace{\left(\sum_{\substack{j=1\\ j\neq k}}^{n} A_{k,j}x_j\right)}_{\text{independent of } x_k} x_k + A_{k,k}x_k^2 + \sum_{i=1}^{n} b_i x_i + c$$

so

$$\frac{\partial f}{\partial x_k}(x) = 2\sum_{\substack{j=1\\ j\neq k}}^{n} A_{k,j}x_j + 2A_{k,k}x_k + b_k = 2\sum_{j=1}^{n} A_{k,j}x_j + b_k.$$

Therefore,

$$\nabla f(x) = \begin{pmatrix} 2\sum_{j=1}^{n} A_{1,j}x_j + b_1 \\ \vdots \\ 2\sum_{j=1}^{n} A_{n,j}x_j + b_n \end{pmatrix} = 2\begin{pmatrix} \sum_{j=1}^{n} A_{1,j}x_j \\ \vdots \\ \sum_{j=1}^{n} A_{n,j}x_j \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = 2Ax + b.$$

(d) **Now suppose we have a vector $y \in \mathbb{R}^n$ and a (tall) matrix $B \in \mathbb{R}^{n\times d}$, i.e., $n > d$. Compute the vector $x \in \mathbb{R}^d$ that minimizes $\|y - Bx\|_2^2$. When is this vector unique?**

Hint: All the portions of the problem so far ought to be useful.

We have

$$\begin{aligned}\|y - Bx\|_2^2 &= (y - Bx)^\top (y - Bx) \\ &= y^\top y - y^\top Bx - x^\top B^\top y + x^\top B^\top Bx \\ &= x^\top \underbrace{B^\top B}_{A} x + \underbrace{-2y^\top B}_{b^\top} x + \underbrace{y^\top y}_{c}\end{aligned}$$

where A is symmetric:

$$A^\top = (B^\top B)^\top = B^\top (B^\top)^\top = B^\top B = A$$

and PSD:

$$x^\top Ax = x^\top B^\top Bx = (Bx)^\top (Bx) = \|Bx\|_2^2 \geq 0.$$

By Question (b): do we have $B^\top y \in \operatorname{ran}(B^\top B)$? is $B^\top B$ invertible?

Interesting case: if $B \in \mathbb{R}^{n \times d}$ has full column rank (which is possible since $n > d$), then $B^\top B \in \mathbb{R}^{d \times d}$ is invertible:

$$B^\top Bx = 0 \Rightarrow x^\top B^\top Bx = 0 \Rightarrow \|Bx\|_2^2 = 0 \Rightarrow Bx = 0 \Rightarrow \sum_{i=1}^{d} x_i B_i = 0$$

where $B_i \in \mathbb{R}^n$ is the i-th column of B. By linear independence of $(B_1, \ldots, B_d)$, $x_1 = \cdots = x_d = 0$, i.e., $x = 0$.

Problem 2 (Probability): Likelihood and posterior. Let X, Y, Z be random vectors of the appropriate dimensions, and suppose we have $Y = AX + Z$ for a matrix A.

(a) **Suppose Z is a standard Gaussian random vector (zero mean, identity covariance). What is the distribution of Y given that $X = x$?**

Clearly, with a fixed value of x, the distribution of Y must be Gaussian with some mean (since it is a fixed value plus a Gaussian).

We have $(Y \mid X = x) = Ax + Z$ so

$$\mathrm{E}_{Y|X=x}[Y] = Ax + \mathrm{E}[Z] = Ax$$

and

$$\begin{aligned}\operatorname{var}_{Y|X=x}[Y] &= \mathrm{E}_{Y|X=x}[(Y - \mathrm{E}_{Y|X=x}[Y])(Y - \mathrm{E}_{Y|X=x}[Y])^\top] \\ &= \mathrm{E}[\big((Ax + Z) - Ax\big)\big((Ax + Z) - Ax\big)^\top] \\ &= \mathrm{E}[ZZ^\top] \\ &= \operatorname{var}[Z] \\ &= I_n\end{aligned}$$

Therefore, $(Y \mid X = x) \sim \mathcal{N}(Ax, I_n)$: a Gaussian random vector with mean Ax and identity covariance.

(b) **Now suppose that we have a *prior* that X is also a standard Gaussian itself, and now we observe the realization $Y = y$. What is the distribution of X given that $Y = y$?**

Hint: Use Bayes' rule.

The key is to understand that the random variables X and Y are themselves jointly Gaussian, and so conditioning on the value of one of them leaves the other Gaussian as well. For a detailed exposition of this, please see pages 268-269 at this link: `http://www.math.chalmers.se/~rootzen/highdimensional/SSP4SE-appA.pdf`. Reading this carefully will also give you an idea of how to derive the mean and covariance of the resulting conditional Gaussian distribution.

We will provide a more detailed solution shortly based on an explicit computation coming from Bayes' rule, stay tuned!