Week 5: Wednesday, Feb 22

Least squares: the big idea

Least squares problems are a special sort of minimization. Suppose $A \in \mathbb{R}^{m \times n}$ and $m > n$. In general, we will not be able to exactly solve *overdetermined* equations $Ax = b$; the best we can do is to minimize the *residual* $r = b - Ax$. In least squares problems, we minimize the two-norm of the residual[1]:

$$\text{Find } \hat{x} \text{ to minimize } \|r\|_2^2 = \langle r, r \rangle.$$

This is not the only way to approximate the solution to an overdetermined system, but it is attractive for several reasons:

1. It's *really* mathematically attractive. $\|x\|^2$ is a smooth function of x, and the solution to the least squares problem is a linear function of b ($x = A^\dagger b$ where $A^\dagger$ is the Moore-Penrose pseudoinverse of A)

2. There's a nice picture that goes with it – the least squares solution is the projection of b onto the span of A, and the residual at the least squares solution is orthogonal to the span of A.

3. It's a mathematically reasonable choice in statistical settings when the data vector b is contaminated by Gaussian noise.

Normal equations

One way to solve the least squares problem is to attack it directly. We know $\|r\|^2 = \|b - Ax\|^2$; and from a given x, the directional derivative in any direction δx is

$$\nabla_x \|r\|^2 \cdot \delta x = 2\langle A\delta x, b - Ax \rangle = 2\delta x^T (A^T b - A^T A x).$$

The minimum occurs when all posible directional derivatives are zero, which gives us the *normal equations*[2]

$$A^T A x = A^T b.$$

[1] Minimizing the two-norm is equivalent to miminizing the squared two-norm.

[2] They are called the *normal* equations because they specify that the residual must be normal (orthogonal) to every vector in the span of A.

Rearranging, we have

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

the matrix $A^\dagger = (A^T A)^{-1} A^T$ is the *Moore-Penrose pseudoinverse* of A (sometimes just called the pseudoinverse).

If the columns of A are not too close to linearly dependent, we would usually just form the normal equations and solve them by using Cholesky factorization to write

$$A^T A = R^T R,$$

where R is an upper triangular matrix.

QR factorization

Another approach is to write a *QR factorization*:

$$A = QR = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1$$

where $Q \in \mathbb{R}^{m \times m}$ is orthogonal ($Q^T Q = I$) and R is upper triangular. The columns of $Q_1 \in \mathbb{R}^{m \times n}$ form an orthonormal basis for the range space of A, and the columns of Q_2 span the orthogonal complement. The factorization $A = Q_1 R_1$ is sometimes called the "economy" QR factorization.

Multiplication by an orthogonal matrix does not change lengths, so

$$\|r\|^2 = \|Q^T r\|^2 = \left\| \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x - Q^T b \right\|^2 = \|R_1 x - Q_1^T b\|^2 + \|Q_2^T b\|^2.$$

The second part of this expression ($\|Q_2^T b\|^2$) is error that we cannot reduce; but $R_1 x - Q_1^T b$ can be made exactly equal to zero. That is, the solution to the least squares problem is

$$x = R_1^{-1} Q_1^T b.$$

In Matlab, we can compute the QR factorization using the `qr` routine:

```
[Q, R ] = qr(A);   % Full QR
[Q1,R1] = qr(A,0); % Economy QR
```

We also use QR implicitly if we solve a least-squares system using the ever-useful backslash operator:

```
x = A\b;  % Minimize norm(Ax−b) via a QR factorization
```

Sensitivity and conditioning

At a high level, there are two pieces to solving a least squares problem:

1. Project b onto the span of A.
2. Solve a linear system so that Ax equals the projected b.

Correspondingly, there are two ways we can get into trouble in solving least squares problem: either b may be nearly orthogonal to the span of A, or the linear system might be ill-conditioned.

Let's consider the issue of b nearly orthogonal to A first. Suppose we have the trivial problem

$$A = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad b = \begin{bmatrix} \epsilon \\ 1 \end{bmatrix}.$$

The solution to this problem is $x = \epsilon$; but the solution for

$$A = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \hat{b} = \begin{bmatrix} -\epsilon \\ 1 \end{bmatrix}.$$

is $\hat{x} = -\epsilon$. Note that $\|\hat{b} - b\|/\|b\| \approx 2\epsilon$ is small, but $|\hat{x} - x|/|x| = 2$ is huge. That is because the projection of b onto the span of A (i.e. the first component of b) is much smaller than b itself; so an error in b that is small relative to the overall size may not be small relative to the size of the projection onto the columns of A.

Of course, the case when b is nearly orthogonal to A often corresponds to a rather silly regression, like trying to fit a straight line to data distributed uniformly around a circle, or trying to find a meaningful signal when the signal to noise ratio is tiny. This is something to be aware of and to watch out for, but it isn't exactly subtle: if $\|r\|/\|b\|$ is close to one, we have a numerical problem, but we also probably don't have a very good model. A more subtle issue problem occurs when some columns of A are nearly linearly dependent (i.e. A is ill-conditioned).

The *condition number of A for least squares* is

$$\kappa(A) = \|A\|\|A^\dagger\| = \kappa(R_1) = \sqrt{\kappa(A^T A)}.$$

We generally recommend solving least squares via QR factorization because $\kappa(R_1) = \kappa(A)$, while forming the normal equations squares the condition number. If $\kappa(A)$ is large, that means:

1. Small relative changes to A can cause large changes to the span of A (i.e. there are some vectors in the span of $\hat{A}$ that form a large angle with all the vectors in the span of A).

2. The linear system to find x in terms of the projection onto A will be ill-conditioned.

If θ is the angle between b and the range of A[3], then the sensitivity to perturbations in b is

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\kappa(A)}{\cos(\theta)} \frac{\|\delta b\|}{\|b\|},$$

while the sensitivity to perturbations in A is

$$\frac{\|\Delta x\|}{\|x\|} \leq \left(\kappa(A)^2 \tan(\theta) + \kappa(A)\right) \frac{\|E\|}{\|A\|}.$$

Even if the residual is moderate, the sensitivity of the least squares problem to perturbations in A (either due to roundoff or due to measurement error) can quickly be dominated by $\kappa(A)^2 \tan(\theta)$ if $\kappa(A)$ is at all large.

In regression problems, the columns of A correspond to explanatory factors. For example, we might try to use height, weight, and age to explain the probability of some disease. In this setting, ill-conditioning happens when the explanatory factors are correlated — for example, perhaps weight might be well predicted by height and age in our sample population. This happens reasonably often. When there is some correlation, we get moderate ill conditioning, and might want to use QR factorization. When there is a lot of correlation and the columns of A are truly linearly dependent (or close enough for numerical work), we have a *rank-deficient* problem. We will talk about rank-deficient problems next lecture.

Problems to Ponder

1. If x minimizes $\|b - Ax\|^2$, argue that $r \perp Ax$.

2. Show that if x is minimizes $\|Ax - b\|$, then $\|Ax\|^2 + \|r\|^2 = \|b\|^2$.

3. Suppose that $A \in \mathbb{R}^{m \times n}$, $m > n$, and that A has full column rank. Then $A^T A$ is symmetric and positive definite. Why?

[3] Note that b, Ax, and r are three sides of a right triangle, so $\sin(\theta) = \|r\|/\|b\|$.

4. Suppose $A^T A = LL^T$, where L is a lower triangular Cholesky factor. Show that the columns of AL^{-T} are orthonormal.

5. Show that minimizing $\|Ax-b\|$ is equivalent to solving the linear system
$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$
Can you think of an advantage of writing the least square problem in this way?

6. Find an orthonormal basis for $\mathcal{P}_2$ with the $L^2([-1,1])$ inner product.

7. How would you find the quadratic $p(x)$ to minimize
$$\int_{-1}^{1} \left(p(x) - f(x)\right)^2 \, dx?$$

8. Suppose $A = \mathbb{R}^{m \times n}$, $m > n$ is full rank, and that $b \in \mathbb{R}^n$. The linear system $A^T x = b$ is *underdetermined.* How would you find the solution that minimizes $\|x\|$?

9. *Maybe only if you've had some stats:* Suppose the entries of $z \in \mathbb{R}^n$ are independent standard normal random variables. Show that for any orthogonal matrix Q, the entries of $Q^T z$ are again independent standard normal random variables.