Cheat sheet: Least Squares Approximation

Viet Duc Nguyen
Technical University of Berlin

December 25, 2018

Contents

0 Preface

A concise overview of the least squares approximation. It is based on the monography *Introduction to Linear Algebra* by Gilbert Strang. This sheet is primarily written for me as a learning guide but may be useful for others. Feel free to use it.

1 Projection onto a line

We consider a line thought the origin. In fact, this case is sufficient, as projecting depends on finding an orthgonal vector, and orthogonality is dependent on the direction of the line and not on its origin.

The projection of an arbitrary point $\mathbf{b} \in \mathbb{R}^d$ on a line G is the point $\mathbf{p}$ on the line G that has the smallest distance to $\mathbf{b}$. That means

$$\mathbf{p} := \min_{\mathbf{g} \in G} \|\mathbf{g} - \mathbf{b}\|.$$

Using geometry, we see that $\mathbf{p}$ is the unique point on G such that $\mathbf{p} - \mathbf{b}$ is orthogonal on G, i.e. the dot-product is zero:

$$\forall \mathbf{g} \in G : (\mathbf{b} - \mathbf{p}) \cdot \mathbf{g} = 0.$$

So, we have two characteerisations of $\mathbf{p}$: (1) the point on the line with the minimal distance to $\mathbf{b}$, and (2) the point on the line such that the vector $\mathbf{p} - \mathbf{b}$ is orthogonal to the line. The second characterisation is useful to compute the point $\mathbf{p}$. We are using the equation:

$$(\mathbf{b} - \mathbf{p})^T \mathbf{g} = 0, \quad \forall \mathbf{g} \in \mathbb{G}.$$

As $\mathbf{p}$ lies on the line $G = \mathbb{R}\mathbf{a}$, there exists $x \in \mathbb{R}$ such that $\mathbf{p} = x\mathbf{a}$. Instead, we are looking for the parameter x, and if we have x, we can obtain $\mathbf{p}$ by

$$\mathbf{p} = x\mathbf{a}.$$

Thus, we solve

$$(\mathbf{b} - x\mathbf{a})^T \cdot \mathbf{a} = 0 \iff \mathbf{b}^T\mathbf{a} - x\mathbf{a}^T\mathbf{a} = 0 \iff x = \frac{\mathbf{b}^T\mathbf{a}}{\mathbf{a}^T\mathbf{a}}.$$

So we get

$$\boxed{x = \frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}}, \quad \mathbf{p} = \frac{\mathbf{a}\mathbf{a}^T}{\mathbf{a}^T\mathbf{a}}\mathbf{b} \quad \text{and} \quad P = \frac{\mathbf{a}\mathbf{a}^T}{\mathbf{a}^T\mathbf{a}}.}$$

P is the projection matrix on G, and the column space of P is G. It holds

$$P^2 = P,$$

since the first projection maps $\mathbf{b}$ onto the column space of P, and thereof, a further projection does nothing.

The perpendicular subspace of P or G is given by

$$G^\perp = I - P.$$

Consider $(I - P)\mathbf{b} = \mathbf{b} - \mathbf{p}$, which is perpendicular to G. The vector $\mathbf{e} = \mathbf{b} - \mathbf{p}$ is also called the *error* vector which becomes clearer in the following sections.

2 Projection onto a subspace

Given a vector $\mathbf{b}$ and a basis $\mathbf{a}_1, ..., \mathbf{a}_n$ find the linear combination of $\mathbf{a}_i$ that is closest to $\mathbf{b}$. This point $\mathbf{p} = A\mathbf{x}$ is called the projection of $\mathbf{b}$ onto the subspace spanned by $\mathbf{a}_i$.

For such point $\mathbf{p}$, the error vector $\mathbf{e} = \mathbf{p} - \mathbf{b}$ must be perpendicular to every basis vector $\mathbf{a}_i$, i.e. for $i = 1, ..., n$:

$$\mathbf{a}_i^T(\mathbf{b} - \mathbf{p}) = 0 \iff \mathbf{a}_i^T(\mathbf{b} - A\mathbf{x}) = 0.$$

This gives

$$A^T(\mathbf{b} - A\mathbf{x}) = 0.$$

We get the famous *least squares equation*:

$$\boxed{A^TA\mathbf{x} = \mathbf{A}^T\mathbf{b}.}$$

The matrix A^TA is $n \times n$ and is symmetric if the columns $\mathbf{a}_i$ are linearly independent. Then the solution for $\mathbf{x}$ becomes

$$\mathbf{x} = (A^TA)^{-1}A^T\mathbf{b}.$$

The projection and projection matrix is given by

$$\boxed{\mathbf{p} = A(A^TA)^{-1}A^T\mathbf{b}, \quad P = A(A^TA)^{-1}A^T}$$

Note that A^TA is indeed invertible, since A consists of linearly independent columns $\mathbf{a}_i$. In fact, A^TA is invertible if and only if it has linearly independent columns, i.e. column rank is full. The matrix A^TA is symmetric, invertible and square, if A has linearly independent columns.

3 Least Squares Approximation

When a matrix A has more rows m than columns n, there might be no solution at all for $A\mathbf{x} = \mathbf{b}$, for A spans a subspace in $\mathbb{R}^m$ with n vectors and $\mathbf{b}$ may not lay in this subspace (due to $m > n$). So we try to find a vector $\mathbf{p} = A\hat{\mathbf{x}}$ that is close to $\mathbf{b}$. As we have seen before, $\mathbf{p}$ is the point in A such that the error vector $p - b \perp A$. To find $\mathbf{p}$, we obtain $\hat{\mathbf{x}}$ by

$$A^TA\hat{\mathbf{x}} = A^T\mathbf{b}.$$

In a idiomatic way, we have just multiplied the matrix A^T to the equation $A\mathbf{x} = \mathbf{b}$ to get a best approximate solution $\hat{\mathbf{x}}$. The error is then given by

$$\mathbf{e} = A\hat{\mathbf{x}} - \mathbf{b}.$$

Another way to see why $\hat{x}$ minimises the $\|\mathbf{e}\|$ is by splitting $\mathbf{b}$ into two parts: $\mathbf{b} = \mathbf{p} + \mathbf{e}$. So $\mathbf{b}$ consists of a vector in A and an orthogonal part $\mathbf{e}$. We cannot solve $A\mathbf{x} = \mathbf{b} = \mathbf{p} + \mathbf{e}$

because $\mathbf{b}$ lies outside of the supspace spanned by A. However, we can solve $A\mathbf{x} = \mathbf{p}$. Thus, the length of any $\mathbf{x}$ in the subspace A to $\mathbf{b}$ is (by theorem of Pythagoras)

$$\|A\mathbf{x} - \mathbf{b}\|^2 = \|A\mathbf{x} - \mathbf{p}\|^2 + \|\mathbf{e}\|^2.$$

We can reduce $\|A\mathbf{x} - \mathbf{p}\|^2$ to zero by choosing $\mathbf{x} = \hat{\mathbf{x}}$, and we are left with the error $\mathbf{e}$, that is unavoidable.

4 Orthogonal bases

When Q consists of orthonormal columns $\mathbf{q}_i$, it holds $Q^TQ = I$ ($m > n$, i.e. full column rank). Such orthogonal matrices also preserve the length of vectors and the dot product, i.e. $(Q\mathbf{x}) \cdot (Q\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$.

Given a least squares problem, the problem of finding the best solution $\hat{\mathbf{x}}$ becomes much more simpler if A is a orthogonal matrix, that we will denote by Q. The formula is then given by

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

and therefore

$$\boxed{\hat{\mathbf{x}} = Q^T\mathbf{b}, \quad \mathbf{p} = QQ^T\mathbf{b}, \quad P = QQ^T.}$$

Note that, $QQ^T \neq I$ if Q is not square. $QQ^T = I$ means that the rows of Q are orthonormal but this cannot be the case if $Q^TQ = I$. The latter requires, that the columns of Q are orthonormal with the number of rows larger than the number of columns $m > n$. So we have m rows with dimension n which cannot be linearly independent, for $m > n$.

If $Q \in \mathbb{R}^{n\times n}$ is square, it spans $\mathbb{R}^n$. The coordinates of a vector $\mathbf{b} \in \mathbb{R}^n$ in respect to the basis $\mathbf{q}_i$ can be easily calculated:

$$\mathbf{b} = (\mathbf{q}_1 \cdot \mathbf{b})\mathbf{q}_1 + ... + (\mathbf{q}_n \cdot \mathbf{b})\mathbf{q}_n.$$

The coordinates are just $\mathbf{x} = Q^T\mathbf{b}$.

5 Gram-Schmidt Process

The idea of the *Gramd-Schmidt process* becomes really simple after all this theory. Starting with a vector $\mathbf{a}_1 = \tilde{\mathbf{a}}_\mathbf{1}$, we orthogonalise a second vector $\mathbf{a}_2$ by just projecting projecting $\mathbf{a}_2$ onto the subspace by $\tilde{\mathbf{a}}_\mathbf{1}$, and take the error vector as the second orthogonalised vector $\tilde{\mathbf{a}}_\mathbf{2}$ (which of course, we have to normalise).

$$\tilde{\mathbf{a}}_\mathbf{2} = \mathbf{a}_2 - \frac{\tilde{\mathbf{a}}_\mathbf{1}\tilde{\mathbf{a}}_\mathbf{1}^T}{\tilde{\mathbf{a}}_\mathbf{1}^T\tilde{\mathbf{a}}_\mathbf{1}}\mathbf{a}_2, \quad \tilde{\mathbf{a}}_\mathbf{3} = \mathbf{a}_3 - \frac{\tilde{\mathbf{a}}_\mathbf{2}\tilde{\mathbf{a}}_\mathbf{2}^T}{\tilde{\mathbf{a}}_\mathbf{2}^T\tilde{\mathbf{a}}_\mathbf{2}}\mathbf{a}_3 - \frac{\tilde{\mathbf{a}}_\mathbf{1}\tilde{\mathbf{a}}_\mathbf{1}^T}{\tilde{\mathbf{a}}_\mathbf{1}^T\tilde{\mathbf{a}}_\mathbf{1}}\mathbf{a}_3.$$