EECS 127/227AT Discussion 3 Slides

Druv Pai

September 19, 2020

Q1

Definition (Least Squares)

An unconstrained optimization problem of the form

$$p^* = \min_x \|y - Ax\|_2^2.$$

Theorem

$x^* = \left(A^\mathsf{T}A\right)^{-1} A^\mathsf{T}y.$

Proof.

Two ways: algebraic or geometric.
Algebraic way: take derivative and set to 0. In particular $\nabla_x \|y - Ax\|_2^2 = 2A^\mathsf{T}Ax - 2A^\mathsf{T}y \overset{\text{set}}{=} 0$; obtains optimal x^*.
Geometric way: error vector $y - Ax^* \perp \text{range}(A)$, since x^* is "best" (use triangle inequality). Thus orthogonal to all columns of A, so $A^\mathsf{T}(y - Ax^*) = 0$, gets same solution. □

Remark: Making updated predictions via least squares has the form: $\widehat{y} = Ax^* = A\left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T}y$.

Remark: Matrix $\left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T}$ is the **left inverse** of A.

Name *left inverse* applies to any matrix A_L^{-1} for which $A_L^{-1}A = I$.
But is there a right inverse?

Q2

SVD: "Generalized diagonalization".
Let $A \in \mathbb{R}^{m \times n}$ and rank$(A) = r$. SVD of A is decomposition

$$A = U\widetilde{\Sigma}V^{\mathsf{T}} = \begin{bmatrix} U_{\mathcal{R}} & U_{\mathcal{N}} \end{bmatrix} \begin{bmatrix} \Sigma & 0^{r \times (n-r)} \\ 0^{(m-r) \times r} & 0^{(m-r) \times (n-r)} \end{bmatrix} \begin{bmatrix} V_{\mathcal{R}}^{\mathsf{T}} \\ V_{\mathcal{N}}^{\mathsf{T}} \end{bmatrix}$$

This is "full SVD". Notice that $A = U_{\mathcal{R}}\Sigma V_{\mathcal{R}}^{\mathsf{T}}$, "compact SVD".

- $U \in \mathbb{R}^{m \times m}$ w/ o.n. columns, $U_{\mathcal{R}} \in \mathbb{R}^{m \times r}$ w/ o.n. columns which span range(A), $U_{\mathcal{N}} \in \mathbb{R}^{m \times (m-r)}$ w/ o.n. columns which span null$\left(A^{\mathsf{T}}\right)$.
- $\widetilde{\Sigma} \in \mathbb{R}^{m \times n}$ diagonal, $\Sigma \in \mathbb{R}^{r \times r}$ diagonal where $\Sigma_{i,i} = \sigma_i = \sqrt{\lambda_i(A^{\mathsf{T}}A)}$ – i^{th} "singular value".
- $V \in \mathbb{R}^{n \times n}$ w/ o.n. columns, $V_{\mathcal{R}} \in \mathbb{R}^{n \times r}$ w/ o.n. columns which span range$\left(A^{\mathsf{T}}\right)$, $V_{\mathcal{N}} \in \mathbb{R}^{n \times (n-r)}$ w/ o.n. columns which span null(A).

Two algorithms to construct SVD:

- Form $V_{\mathcal{R}}$ from eigenvector basis of $A^{\mathsf{T}}A$ and fill Σ with square roots of corresponding eigenvalues.
- For i^{th} column u_i of $U_{\mathcal{R}}$, set $u_i = \frac{1}{\sigma_i}Av_i$ (v_i is i^{th} column of $V_{\mathcal{R}}$).
- Fill up $U_{\mathcal{N}}, V_{\mathcal{N}}$ by picking any basis for $\mathbb{R}^m, \mathbb{R}^n$ that include columns of $U_{\mathcal{R}}, V_{\mathcal{R}}$ and using Gram-Schmidt process

Or: do the same thing except fill up $U_{\mathcal{R}}$ first by using eigenvector basis of AA^{T} and filling in $\sigma_i = \sqrt{\lambda_i(AA^{\mathsf{T}})}$. Do the same process except swapping U and V.
Why? Sometimes AA^{T} or $A^{\mathsf{T}}A$ is a lot easier to compute/smaller.
Why are these constructions equal/justified? Symmetric matrices $A^{\mathsf{T}}A$, AA^{T}, diagonalized as:

$$A^{\mathsf{T}}A = \left(U\widetilde{\Sigma}V^{\mathsf{T}}\right)^{\mathsf{T}}\left(U\widetilde{\Sigma}V^{\mathsf{T}}\right) = V\widetilde{\Sigma}^{\mathsf{T}}U^{\mathsf{T}}U\widetilde{\Sigma}^{\mathsf{T}}V^{\mathsf{T}} = V\widetilde{\Sigma}^{\mathsf{T}}\widetilde{\Sigma}V^{\mathsf{T}}.$$

$$AA^{\mathsf{T}} = \left(U\widetilde{\Sigma}V^{\mathsf{T}}\right)\left(U\widetilde{\Sigma}V^{\mathsf{T}}\right)^{\mathsf{T}} = U\widetilde{\Sigma}V^{\mathsf{T}}V\widetilde{\Sigma}^{\mathsf{T}}U^{\mathsf{T}} = U\widetilde{\Sigma}\widetilde{\Sigma}^{\mathsf{T}}U^{\mathsf{T}}.$$

Pattern matching the diagonalization gives the construction.

If $A = U_{\mathcal{R}} \Sigma V_{\mathcal{R}}^{\mathsf{T}}$ (compact SVD), then Moore-Penrose psuedoinverse is given by $A^{\dagger} = V_{\mathcal{R}} \Sigma^{-1} U_{\mathcal{R}}^{\mathsf{T}}$.
We want to show that $A^{\dagger} y = x^*$ gives the optimal solution to the constrained optimization problem

$$p^* = \min_x \|x\|_2^2 \quad \text{s.t. } Ax = y.$$

This is **least norm** problem.
When A has full column rank (linearly independent columns) then $A^{\dagger} = \left(A^{\mathsf{T}} A\right)^{-1} A^{\mathsf{T}}$, is the **left inverse**.
When A has full row rank (linearly independent rows), then $A^{\dagger} = A^{\mathsf{T}} \left(A A^{\mathsf{T}}\right)^{-1}$, is the **right inverse**.

Q3

Definition

Let $X \in \mathbb{R}^{m \times n}$ be data matrix; **columns are data points, rows are features**. Assume sum of columns is 0 (X is **centered**). Then $\mathrm{Var}(X) = \frac{1}{n} X X^{\mathsf{T}} \in \mathbb{R}^{m \times m}$ is the sample (empirical) variance-covariance matrix of the features.

Important! Most of the time this is flipped around, and you have to take transposes.
What this means is that $u^{\mathsf{T}} C u$ is the variance of the sample data along direction u. Covariance matrix is aligned with coordinate axes; u is our axis to compute covariance along.

Definition

PCA: eigendecomposition of the (symmetric) covariance matrix. Eigenvalues λ_i determine covariance along direction of eigenvector v_i. We can pick a few eigenvectors with largest eigenvalues and replace our data set by the projections onto the space spanned by the v_i; saves a lot of data.