annotated version

Machine Learning Course - CS-433

Least Squares

Oct 4, 2016

minor changes by Martin Jaggi 2016

Motivation

In rare cases, one can compute the optimum of the cost function analytically. Linear regression using MSE is one such case. Here the solution can be obtained explicitly, from using a linear equation system, which is called the normal equations. This method is one of the most popular methods for data fitting, and called least squares.

To derive the normal equations, we use the optimality conditions for convex functions. See the previous lecture notes on optimization.

$$\nabla \mathcal{L}(\mathbf{w}^\star) \stackrel{!}{=} \mathbf{0}$$

This is a system of D equations.

Exercise: Derive the normal equation for a 1-parameter linear model, for $\mathcal{L} = MSE$.

find w

$$MSE(w) = \frac{1}{2N} \sum_{n=1}^{N} (y_n - x_n^T w)^2$$

1-parameter model

$$L(w_0) = \frac{1}{N} \sum \frac{1}{2}(y_n - w_0)^2$$

$$\frac{\partial L}{\partial w_0} = \frac{1}{N} \sum (y_n - w_0)(-1)$$

$$= w_0 - \frac{1}{N} \sum y_n$$

$$\stackrel{!}{=} 0$$

$$\Rightarrow w_0 = \frac{1}{N} \sum y_n = \bar{y}$$

$\mathcal{L}$ = MSE

Normal Equations

Recall the expression of the gradient for multiple linear regression, under mean squared error:

$$\nabla \mathcal{L}(\mathbf{w}) = -\frac{1}{N}\mathbf{X}^\top \mathbf{e} = -\frac{1}{N}\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\mathbf{w})$$

$\mathbf{e}$

$\in \mathbb{R}^D$

Set it to zero to get the normal equations for linear regression.

$$\mathbf{X}^\top \mathbf{e} = \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) \stackrel{!}{=} \mathbf{0}$$

D equations

D = # parameters

columns of X

implying that the error is orthogonal to all rows of $\mathbf{X}^\top$ (= columns of $\mathbf{X}$).

$\mathcal{L}$ convex :

$\nabla \mathcal{L} = 0$ „$\Leftarrow$" optimality

$\Leftarrow$

Geometric Interpretation

$y \approx Xw$ (error)

Denote the d'th column of $\mathbf{X}$ by $\mathbf{x}_{:d}$.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1D} \\ x_{21} & x_{22} & \dots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{ND} \end{bmatrix}$$

$x_{:d}$ (column), x_{1}, x_{n}, x_{N} (rows)

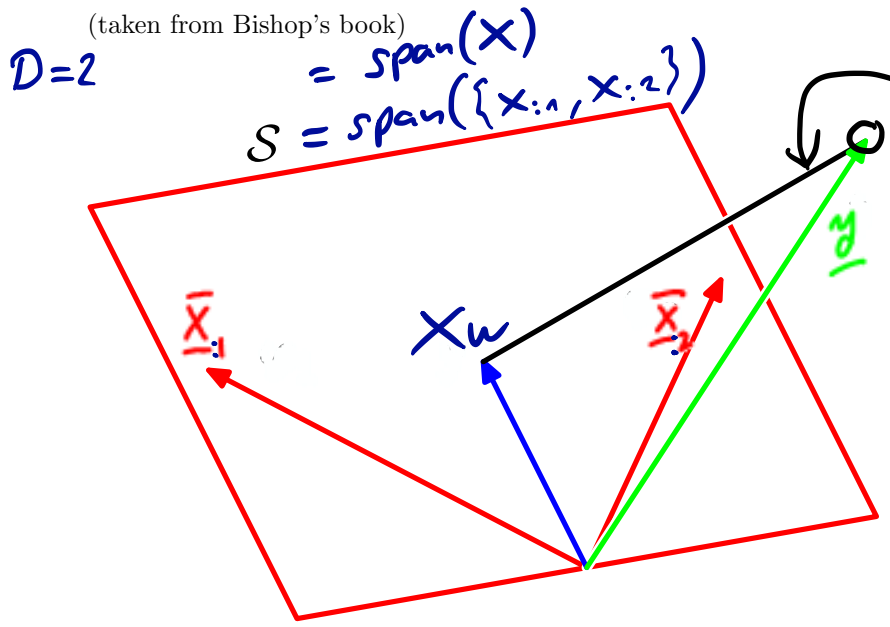
The normal equations tell us to choose a vector in the span of $\mathbf{X}$, such that the error vector $\mathbf{e}$ will be orthogonal to the span.

$\in \mathbb{R}^N$

$\text{span}(\{u, v\}) = \{w_1 u + w_2 v \mid w_1 \in \mathbb{R}, w_2 \in \mathbb{R}\}$

The following figure illustrates this:

(taken from Bishop's book)



Normal Equations

$X^T(y - Xw) \doteq 0$

$Xw = w_1 x_{:1} + w_2 x_{:2}$

Q1: what if $y \in \text{span}(X)$?

Q2: what if $x_{:1} = x_{:2}$?

Q3: what if $y \perp \text{span}(X)$?

$(\Rightarrow Xw = 0)$

Least Squares

When $\mathbf{X}^\top\mathbf{X}$ is invertible, we have a closed-form expression for the minimum.

$$\mathbf{w}^\star = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

We can use this model to predict a new value for an unseen datapoint (test point) $\mathbf{x}_m$:

$$\hat{y}_m := \mathbf{x}_m^\top\mathbf{w}^\star = \mathbf{x}_m^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

find w:

$$X^\top(y - Xw) = 0$$

$$X^\top y = X^\top X w$$

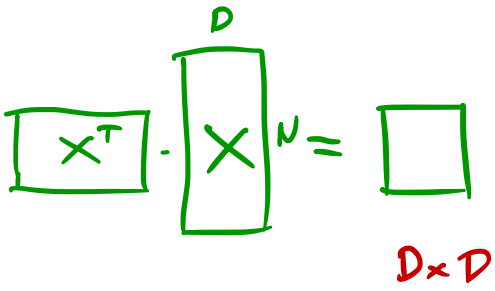
$X^\top X \in \mathbb{R}^{D\times D}$

solve a linear system

w^*

Invertibility and Uniqueness

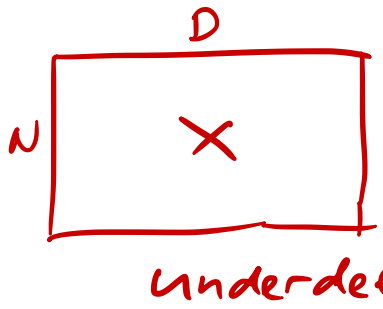
The Gram matrix $\mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{D\times D}$ is invertible iff $\mathbf{X}$ has full column rank, or in other words $rank(\mathbf{X}) = D$.



Proof: Fundamental theorem of linear algebra.

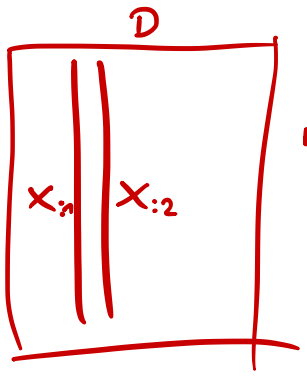
Rank Deficiency and Ill-Conditioning

Unfortunately, in practice, $\mathbf{X}$ is often rank deficient! (meaning $rank(\mathbf{X}) < D$)



underdetermined

- if $D > N$, we always have $rank(\mathbf{X}) < D$
  (since row rank = col. rank)



rank < D

- if $D \leq N$, but some of the columns $\mathbf{x}_{:d}$ are (nearly) collinear. In the later case, the matrix is ill-conditioned, leading to numerical issues when solving the linear system.

condition number

$$\frac{\lambda_{max}(X^TX)}{\lambda_{min}(X^TX)}$$

Summary of Linear Regression

We have studied three types of methods:

1. Grid Search
2. Iterative Optimization Algorithms
   (Stochastic) Gradient Descent
3. Least squares
   closed-form solution, for linear MSE

$$(X^TX)^{-1}X^Ty$$

ND^2 comp. gram + D^3 comp $^{-1}$ + $N \cdot D$ + D^2

computational cost?

$$O(ND^2 + D^3)$$

Additional Notes

Closed-form solution for MAE

Can you derive closed-form solution for 1-parameter model when using MAE cost function?
See this short article: http://www.johnmyleswhite.com/notebook/2013/03/22/modes-medians-and-means-an-unifying-perspective/.

Implementation

There are many ways to solve a linear system, but using the QR decomposition is one of the most robust ways. Matlab's backslash operator and also NumPy's linalg package implement this in just one line:

```
w = np.linalg.solve(A, b)
```

For a robust implementation, see Sec. 7.5.2 of Kevin Murphy's book.

ToDo

1. Revise linear algebra to understand why $\mathbf{X}$ needs to have full rank. Read the Wikipedia page on the rank of a matrix.
2. Understand the robust implementation of the linear system solver, and play with it during the lab. Read Kevin Murphy's Section 7.5.2 for details.
3. Understand ill-conditioning. Reading about the "condition number" in Wikipedia will help. Also, understanding SVD is essential. Here is another link provided by Dana Kianfar (EPFL) http://www.cs.uleth.ca/~holzmann/notes/illconditioned.pdf.
4. Work out the computational complexity of least-squares, when using an existing linear systems solver (use the Wikipedia page on computational complexity).