Linear regression unit:

- Simple linear regression with ordinary least squares (Monday)
- Lab activity: Linear regression (Wednesday)
- Deriving a closed form solution (**today**)
- Newton's method and gradient descent (next week Monday)
- Training linear regression using gradient descent (next week Wednesday)

Today:

- Deriving simple linear regression
- Deriving multiple linear regression
- Deriving MLR with ridge or LASSO regularization

Simple linear regression:

$$y(x) = \theta_0 + \theta_1 x$$

Loss function (sum square error):

$$\mathcal{L}(\vec{\theta}) = \sum_{n=0}^{N-1} (y_n - y(x_n))^2 = \sum_{n=0}^{N-1} (y_n - \theta_0 - \theta_1 x_n)^2$$

Partial derivatives of the loss function:

$$\frac{\partial \mathcal{L}}{\partial \theta_0} = -2 \sum_{n=0}^{N-1} (y_n - \theta_1 x_n - \theta_0) \qquad \frac{\partial \mathcal{L}}{\partial \theta_1} = \sum_{n=0}^{N-1} -2x_n(y_n - \theta_1 x_n - \theta_0)$$

Closed form solution:

$$\theta_0 = \bar{y} - \theta_1 \bar{x} \qquad \theta_1 = \frac{\sum_{n=0}^{N-1} (x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=0}^{N-1} (x_n - \bar{x})^2}$$

... where $\bar{y}$ and $\bar{x}$ are the mean values of y and x

Multiple linear regression:

$$y(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_D x_D = \theta_0 + \boldsymbol{\theta}^T \boldsymbol{x}$$

Most general form of linear regression on arbitrary basis functions $\phi_1 \ldots \phi_D$:

$$y(\boldsymbol{x}) = \theta_0 + \theta_1 \phi_1(\boldsymbol{x}) + \cdots \phi_D(\boldsymbol{x})$$

Polynomial regression—assume original data is scalar and basis functions are $\phi_i(x) = x^i$.

$$y(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots \theta_D x^D$$

(It's called *linear regression* because the components are combined linearly.)

Multiple linear regression:

$$y(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_D x_D = \theta_0 + \boldsymbol{\theta}^T \boldsymbol{x}$$

If we extend each observation so that it has 1 in position 0, that is $\boldsymbol{x} = [1, x_1, x_2, \ldots x_D]$ (so each observation acts like a vector of length $D + 1$), and interpret $\boldsymbol{\theta}$ as $[\theta_0, \theta_1, \theta_2, \ldots \theta_D]$, then the model family is

$$y(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x}$$

$$y(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x}$$

Loss function:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \sum_{n=0}^{N-1} (y_n - y(\boldsymbol{x}_n))^2 \\
&= \sum_{n=0}^{N-1} (y_n - \theta_0 - \theta_1 x_{n,1} \cdots - \theta_D x_{n,D})^2 \\
&= ||\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta}||_2^2 && L_2 \text{ (Euclidean) norm, squared} \\
&= (\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta})^T (\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta}) && \text{“linear algebra” form}
\end{aligned}$$

Partial derivatives of the loss function, "non-linear-algebra form":

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=0}^{N-1}(y_n - \theta_0 - \theta_1 \boldsymbol{x}_{n,1} \cdots - \theta_D \boldsymbol{x}_{n,D})^2$$

$$\frac{\partial \mathcal{L}}{\partial \theta_0} = -2 \sum_{n=0}^{N-1}(y_n - \theta_0 - \theta_1 \boldsymbol{x}_{n,1} \cdots - \theta_D \boldsymbol{x}_{n,D})$$

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = -2 \sum_{n=0}^{N-1} \boldsymbol{x}_{n,i}(y_n - \theta_0 - \theta_1 \boldsymbol{x}_{n,1} \cdots - \theta_D \boldsymbol{x}_{n,D})$$

Redone in "linear-algebra form":

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \sum_{n=0}^{N-1}(y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2 \\
&= (\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta}) \\
&= (\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta})
\end{aligned}
$$

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}}(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}) \\
&= -2\boldsymbol{y}^T\mathbf{X} + 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}
\end{aligned}
$$

Now we set the whole lot of the partial derivatives to $\mathbf{0}$, that is, the zero vector of length $D + 1$, and solve for $\boldsymbol{\theta}$.

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) &= -2\boldsymbol{y}^T \mathbf{X} + 2\boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \\
\mathbf{0} &= -2\boldsymbol{y}^T \mathbf{X} + 2\boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \\
\boldsymbol{y}^T \mathbf{X} &= \boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \\
\boldsymbol{\theta}^T &= \boldsymbol{y}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\
\boldsymbol{\theta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{y}
\end{aligned}
$$

5.1.2 Explicit solution

Least squares is the method that solves the empirical risk minimization problem for the hypothesis class (5.1) with respect to the squared loss. We want to find w that minimizes

$$\arg\min_w C(w) = \arg\min_w L(f_w) = \arg\min_w \frac{1}{2m}\sum_{i=1}^{m}(w^T x_i - y_i)^2.$$

Note that here we use the homogeneous notation: $w = (w_1, \ldots, w_n, b)$, $x_i = (x_{i1}, \ldots, x_{in}, 1)^T$. We will use the more compact notation and equivalent formulation

$$\arg\min_w C(w) = \frac{1}{2m}\arg\min_w \|Xw - Y\|^2, \tag{5.3}$$

where $X = (x_{ij})_{ij} \in \mathbb{R}^{m\times n}$, $Y = (y_1, \ldots, y_m)^T$, and $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^m$. The number m is the number of samples, and n is the number of *features*.

Han Veiga and Ged, pg 72

Corollary 5.1.4. *Following Theorem 5.1.3 and assuming that the data $\{x_1, \ldots, x_m\}$ are not colinear, we can specify some properties of the solution w:*

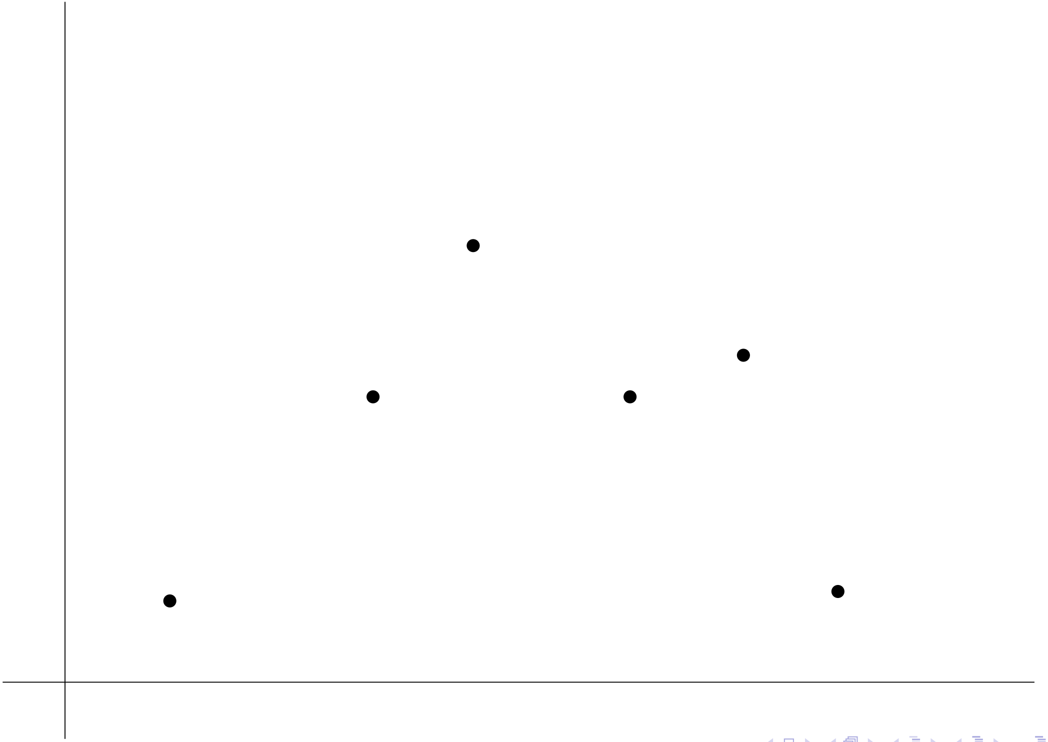
(i) *When $n = m$, we have by definition $X^+ = X^{-1}$ and thus $w = X^{-1}Y$.*

(ii) *When $m > n$, X^TX is invertible, and there is a unique $w = (X^TX)^{-1}X^TY$.*

(iii) *When $n > m$, X^TX is not invertible, and there are infinitely many solutions w.*

The proof of this corollary is left as an exercise to the reader.

Han Veiga and Ged, pg 74



Loss function for ridge regularization (ridge regression):

$$\mathcal{L}_{ridge}(\boldsymbol{\theta}) = \underbrace{||\boldsymbol{y}^T - \boldsymbol{\theta}^T\mathbf{X}||_2^2}_{\text{original loss}} + \underbrace{\alpha||\boldsymbol{\theta}||_2^2}_{\text{regularizer}}$$

Finding a closed form for ridge regression:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) &= -2\boldsymbol{y}^T\mathbf{X} + 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X} + 2\alpha\boldsymbol{\theta} \\
\mathbf{0} &= -2\boldsymbol{y}^T\mathbf{X} + 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X} + 2\alpha\boldsymbol{\theta} \\
\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X} + \alpha\boldsymbol{\theta} &= \boldsymbol{y}^T\mathbf{X} \\
\boldsymbol{\theta}^T(\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I}) &= \boldsymbol{y}^T\mathbf{X} \\
\boldsymbol{\theta}^T &= \boldsymbol{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1} \\
&= (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{y}
\end{aligned}$$

Loss function for ridge regularization:

$$\mathcal{L}_{ridge}(\boldsymbol{\theta}) = \underbrace{||\boldsymbol{y}^T - \boldsymbol{\theta}^T \mathbf{X}||_2^2}_{\text{original loss}} + \underbrace{\alpha||\boldsymbol{\theta}||_2^2}_{\text{regularizer}}$$

Loss function for LASSO regularization

$$\begin{aligned}\mathcal{L}_{LASSO}(\boldsymbol{\theta}) &= ||\boldsymbol{y}^T - \boldsymbol{\theta}^T \mathbf{X}||_2^2 + \alpha||\boldsymbol{\theta}||_1 \\ &= ||\boldsymbol{y}^T - \boldsymbol{\theta}^T \mathbf{X}||_2^2 + \alpha \textstyle\sum_{i=1}^{D} |\theta_i|\end{aligned}$$

Coming up:

Due Thurs, Jan 30:
Read the textbook from Chapters 1 and 5 (see Canvas for specific sections)

Due Fri, Jan 31:
Do KNN programming assignment

Due Tues, Feb 4:
Take linear regression quiz
Propose project topic

Due Thurs, Feb 6:
Read textbook from Chapter 3 (see Canvas for details)

Due Fri, Feb 7:
Do linear regression programming assignment