Linear regression unit:

- Linear regression general concepts (last week Monday)
- Lab activity: Linear regression (last week Wednesday)
- Deriving an explicit solution to linear regression (last week Friday)
- Regularization; Newton's method (Monday)
- Gradient descent and its use in linear regression (**today**)

Today:

- Quiz questions
- Introducing gradient descent
- The cost of linear regression
- A gradient-descent solution to linear regression

What makes linear regression *linear*?

- It finds the line of best fit.
- You use linear algebra to do it.
- Each term is a linear function of one or more of the original features.
- The (original or computed) features are combined linearly.
- It was invented by Carl Linnaeus.
- It was invented by Linus Torvalds.
- It was invented by Linus Pauling.
- It was named in honor of Linus VanPelt.

How does multiple regression differ from simple linear regression?

- It does linear regression multiple times.
- It does simple regression on multiple lines.
- It has no closed form solution.
- It does linear regression on higher dimensional data.

Which is **not** true of regularization?

- It is used to counteract overfitting.
- It works by penalizing model complexity.
- It works by reducing the influence of less-informative variables.
- It is an example of a normal equation.

Match **Ridge** and **LASSO** each with the norm it uses in its penalty term.

- L1 (Manhattan)
- L2 (Euclidean)
- Mahalanobis
- Canberra

Linear regression (unregularized)

Closed form solution (using *sum square error* as loss function):

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$$

A new loss function, *mean square error*:

$$\mathcal{L}_{MSE}(\boldsymbol{\theta}) = \frac{1}{N}||\boldsymbol{y}^T - \boldsymbol{\theta}^T\mathbf{X}^T||^2 = \frac{1}{N}||\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta}||^2$$

The gradient of this loss function:

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}(-2\boldsymbol{y}^T\mathbf{X} + 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X})$$

Ridge regression

Closed form solution (using *sum square error* as loss function):

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X} + \mathbf{A})^{-1}\mathbf{X}^T\boldsymbol{y}$$

Mean square error loss and gradient:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{N}||\boldsymbol{y}^T - \boldsymbol{\theta}^T\mathbf{X}||^2 + \alpha||\boldsymbol{\theta}||_2^2 \\
\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{N}(-2\boldsymbol{y}^T\mathbf{X} + 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}) + 2\alpha\boldsymbol{\theta}
\end{aligned}$$

LASSO

(No closed form solution)

Mean square error loss and subgradient:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{N}||\boldsymbol{y}^T - \boldsymbol{\theta}^T\mathbf{X}||^2 + \alpha||\boldsymbol{\theta}||_1 \\ \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{N}(-2\boldsymbol{y}^T\mathbf{X} + 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}) + \alpha\,(\text{sign}(\boldsymbol{\theta}))\end{aligned}$$

Where

$$\text{sign}(x) = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases} \qquad \text{and} \qquad \text{sign}(\boldsymbol{x}) = [\text{sign}(x_0)\ldots\text{sign}(x_{D-1})]$$

Coming up:

Due Thurs, Feb 6:
Read textbook from Chapter 3 (see Canvas for details)
Take gradient descent and linear regression quiz

Due Fri, Feb 7:
Do linear regression programming assignment

Due Wed, Feb 12:
Do gradient descent for linear regression programming assignment