Linear regression unit:

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

Today:

- Tidying up recent loose ends
- Newton's method and a sample iterative method
- The gradient descent algorithm

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.

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

Note that $\sum_{n=0}^{N-1}(\bar{y}-y_n)=0$ and $\sum_{n=0}^{N-1}(\bar{x}-x_n)=0$, and so $\sum_{n=0}^{N-1}\bar{x}(\bar{y}-y_n)=0$ and $\sum_{n=0}^{N-1}\bar{x}(\bar{y}-y_n)=0$. Plugging these in...

$$\begin{aligned}
\theta_1 &= \frac{\sum_{n=0}^{N-1} x_n(y_n-\bar{y})}{\sum_{n=0}^{N-1} x_n(x_n-\bar{x})} \\
&= \frac{\sum_{n=0}^{N-1} x_n(y_n-\bar{y}) + \sum_{n=0}^{N-1}\bar{x}(\bar{y}-y_n)}{\sum_{n=0}^{N-1} x_n(x_n-\bar{x}) + \sum_{n=0}^{N-1}\bar{x}(\bar{x}-x_n)} \\
&= \frac{\sum_{n=0}^{N-1}(x_n-\bar{x})(y_n-\bar{y})}{\sum_{n=0}^{N-1}(x_n-\bar{x})^2}
\end{aligned}$$

Summary of **simple linear regression** using **least squares**:

Let $\bar{x}$ and $\bar{y}$ be the mean observation and target values, respectively. Then the line of best fit is

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

where

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

Root mean square error:

$$\mathcal{L}_{RMSE}(\boldsymbol{\theta}) = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}(y_n - y(\boldsymbol{x}_n))^2}$$

Sum square error:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=0}^{N-1}(y_n - y(\boldsymbol{x}_n))^2$$

Sum square error, ‘linear-algebra form”:

$$\mathcal{L}(\boldsymbol{\theta}) = ||\boldsymbol{y}^T - \mathbf{X}\boldsymbol{\theta}||^2$$

Partial derivatives of the sum square error, "non-linear-algebra form":

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

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

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

Redone in "linear-algebra form":

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

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

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 $\boldsymbol{\theta}$.

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

Loss function for ridge regularization (ridge regression):

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

Finding a closed form for ridge regression (almost):

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

Loss function for LASSO regularization

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

Loss function for ridge regularization done more carefully:

$$\mathcal{L}_{ridge}(\boldsymbol{\theta}) = \underbrace{||\boldsymbol{y}^T - \boldsymbol{\theta}^T \mathbf{X}||^2}_{\text{original loss}} + \underbrace{\alpha \sum_{i=1}^{D} \theta_i^2}_{\text{regularizer}}$$

Finding a closed form for ridge regression. Let $\hat{\boldsymbol{\theta}}$ be $\boldsymbol{\theta}$ but with 0 in index 0.

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

where $\mathbf{A}$ is like αI but with 0 in the top left corner.

Deriving Newton's method: Suppose we have a function f with derivative f'. (If we don't know f' then we can approximate it numerically.) We want to find a root x_r, that is an x value where the curve of f crosses the x-axis, $f(x_r) = 0$.
Let x_0 be a guess at the root. Then

$$
\begin{aligned}
y - y_0 &= m(x - x_0) \\
y - f(x_0) &= f'(x_0)(x - x_0) \\
y &= f'(x_0)(x - x_0) + f(x_0) \\
y &= f'(x_0)x + (f(x_0) - x_0 f'(x_0))
\end{aligned}
$$

Set $y = 0$ for this tangent and solve for x.

$$
\begin{aligned}
0 &= f'(x_0)x + (f(x_0) - x_0 f'(x_0)) \\
f'(x_0)x &= x_0 f'(x_0) - f(x_0) \\
x_1 &= \tfrac{x_0 f'(x_0) - f(x_0)}{f'(x_0)}
\end{aligned}
$$

To compute an improved guess x_{i+1} over a current guess x_i:

$$x_{i+1} = \frac{x_i f'(x_i) - f(x_i)}{f'(x_i)}$$

Coming up:

Read textbook sections on linear regression(due end-of-day Mon, Jan 30)
Do linear regression assignment (due end-of-day Tues, Jan 31)

Take gradient descent quiz (due classtime Fri, Feb 3)

Project proposal *(due end-of-day Fri, Feb 3)*