Linear regression unit:

- Simple linear regression with ordinary least squares (**today**t)
- Lab activity: Linear regression (Wednesday)
- Newton's method and gradient descent (Friday)
- Training linear regression using gradient descent (next week Monday)

Today:

- Foundational ideas
  - Problem statement for linear regression
  - Error and loss
  - Partial derivatives and gradients
- Deriving ordinary least squares for simple linear regression
- Multiple regression
- Regularization

Which of the following is *not* a hyperparameter of a k nearest neighbor classifier?

The number of neighbors   The distance metric   The curse of dimensionality

Which of the following is *not* true about k nearest neighbors classification?

It is non-parametric   The classifier stores all the training data

It works very well on many applications   The curse of dimensionality doesn't apply

Which if the following is *not* a vector distance metric?

Euclidean   Manhattan   Mahalanobis

Canberra   Curse of dimensionality

Finding the partial derivative of the loss function with respect to θ_0 and finding θ_0 where this partial derivative is 0:

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

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

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

Doing the same but for θ_1:

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

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

Note that $\sum_{n=0}^{N-1}(\bar{y} - y_n) = 0$ and $\sum_{n=0}^{N-1}(\bar{y} - y_n) = 0$, 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...

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

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":

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

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 $\boldsymbol{\theta}$.

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

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:

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

Loss function for LASSO regularization

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

Coming up:

Read and respond to Boston Housing Dataset article (see Schoology for details) (due Tuesday)

Do K nearest neighbors assignment (due end-of-day Thurs)

Take linear regression quiz (due class time Fri)

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