Logistic regression unit:

- Derivation from linear regression (**today**)
- Lab activity: Applying logistic regression (next week Monday)
- Multiclass classification (next week Wednesday)

Today:

- Adapting regression to classification
- Evaluating classification
- Introducing the logistic function

Courses that cover Newton's method:

MATH 231

CSCI 243

CSCI 381

Things Newton's method and gradient descent have in common:

They both involve following a slope to improve a guess

They both follow the algorithmic pattern of improve-until-good-enough.

They both involve making an initial guess.

They both have an intuitive geometric interpretation.

Reasons to use gradient descent:

There is no closed form solution for finding the optimal parameters for the model.

Computing the optimal parameters for the model directly is too expensive.

The number of points in the training data is too large to fit into memory at once.

Possible termination conditions for gradient descent:

- *The norm of the gradient is within a tolerance*
- *The improvement of the loss function between successive iterations is less than a tolerance*
- *The maximum number of iterations has been reached.*

Parameters to or variations of gradient descent:

- *The learning rate*
- *The termination condition*
- *The function to minimize, together with its gradient*
- *Batch vs stochastic vs minibatch*
- *The maximum number of iterations*

Let N be the number of data points and D be the dimensionality of the data. Then

$\mathbf{X}$	$N \times D$	$\boldsymbol{y}^T\mathbf{X}$	$1 \times D$
$\boldsymbol{y}$	$N \times 1$	$\mathbf{X}^T\mathbf{X}$	$D \times D$
$\boldsymbol{\theta}$	$D \times 1$	$\boldsymbol{\theta}^T\mathbf{X}^T$	$1 \times N$
$\mathcal{L}(\boldsymbol{\theta})$	scalar	$\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}$	$1 \times D$
$\nabla_{\boldsymbol{\theta}}\mathcal{L}$	$D \times 1$		

To compute $||\boldsymbol{y}^T - \boldsymbol{\theta}^T \mathbf{X}^T||^2$, each part costs

$\boldsymbol{\theta}^T \mathbf{X}^T$	$O(ND)$
Vector subtraction	$O(N)$
Norm	$O(N)$
Dominant term	$O(ND)$

Opportunities of the **Logistic Regression** topic:

- How to adapt regression to classification.
- How to measure classification performance using accuracy, precision, and recall.
- The nifty mathematics of the logistic function.
- How to use probabilities for classification.
- How to adapt binary classification to multiway classification.
- The mathematics of the softmax function.
- A reinforcement of gradient descent.

Measurements to evaluate a (binary) classifier (based on true positives TP, true negatives TN, false positives FP, false negatives FN):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

Coming up:

Due Fri, Feb 7:
Do linear regression programming assignment

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

(There will be a logistic regression quiz, probably due Thurs, Feb 13)

Due Wed, Feb 19:
Do logistic regression programming assignment

Due Fri, Feb 21:
Submit "Dataset" checkpoint for term project