

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

$$\mathbf{X} \quad N \times D$$

$$\mathbf{y}^T \mathbf{X} \quad 1 \times D$$

$$\mathbf{y} \quad N \times 1$$

$$\mathbf{X}^T \mathbf{X} \quad D \times D$$

$$\boldsymbol{\theta} \quad D \times 1$$

$$\boldsymbol{\theta}^T \mathbf{X}^T \quad 1 \times N$$

$$\mathcal{L}(\boldsymbol{\theta}) \quad \text{scalar}$$

$$\boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \quad 1 \times D$$

$$\nabla_{\boldsymbol{\theta}} \mathcal{L} \quad D \times 1$$

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

$$\boldsymbol{\theta}^T \mathbf{X}^T \quad O(ND)$$

$$\text{Vector subtraction} \quad O(N)$$

$$\text{Norm} \quad O(N)$$

$$\text{Dominant term} \quad O(ND)$$

Coming up:

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

Do gradient descent/linear regression assignment (due end-of-day Tues, Feb 7)

Take logistic regression quiz (due classtime Wed, Feb 8)