Classification
Logistic Regression

UNIVERSITY *of* WASHINGTON

Loss function: Conditional Likelihood

- **Have a bunch of iid data:** $\{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^d, \quad y_i \in \{-1, 1\}$

$$P(Y = -1|x, w) = \frac{1}{1 + \exp(w^T x)}$$

$$P(Y = 1|x, w) = \frac{\exp(w^T x)}{1 + \exp(w^T x)}$$

- **This is equivalent to:**

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y\, w^T x)}$$

- **So we can compute the maximum likelihood estimator:**

$$\boxed{\widehat{w}_{MLE} = \arg\max_w \prod_{i=1}^n P(y_i|x_i, w)}$$

Sigmoid for binary classes

$$\mathbb{P}(Y=0|w,X) = \boxed{\frac{1}{1+\exp(w_0+\sum_k w_k X_k)}}$$

$$\mathbb{P}(Y=1|w,X) = 1-\mathbb{P}(Y=0|w,X) = \frac{\exp(w_0+\sum_k w_k X_k)}{1+\exp(w_0+\sum_k w_k X_k)}$$

$$\log\left\{\frac{\mathbb{P}(Y=1|w,X)}{\mathbb{P}(Y=\cancel{0}_{-1}|w,X)}\right\} = \exp(w_0+\sum_k w_k X_k)$$

$\geq 1 \rightarrow Y=1$ is more likely

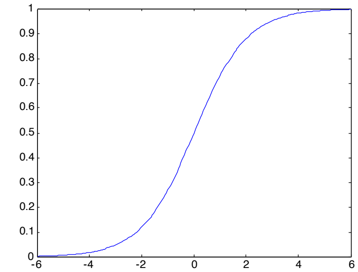
$< 1 \rightarrow Y=-1$ is more likely

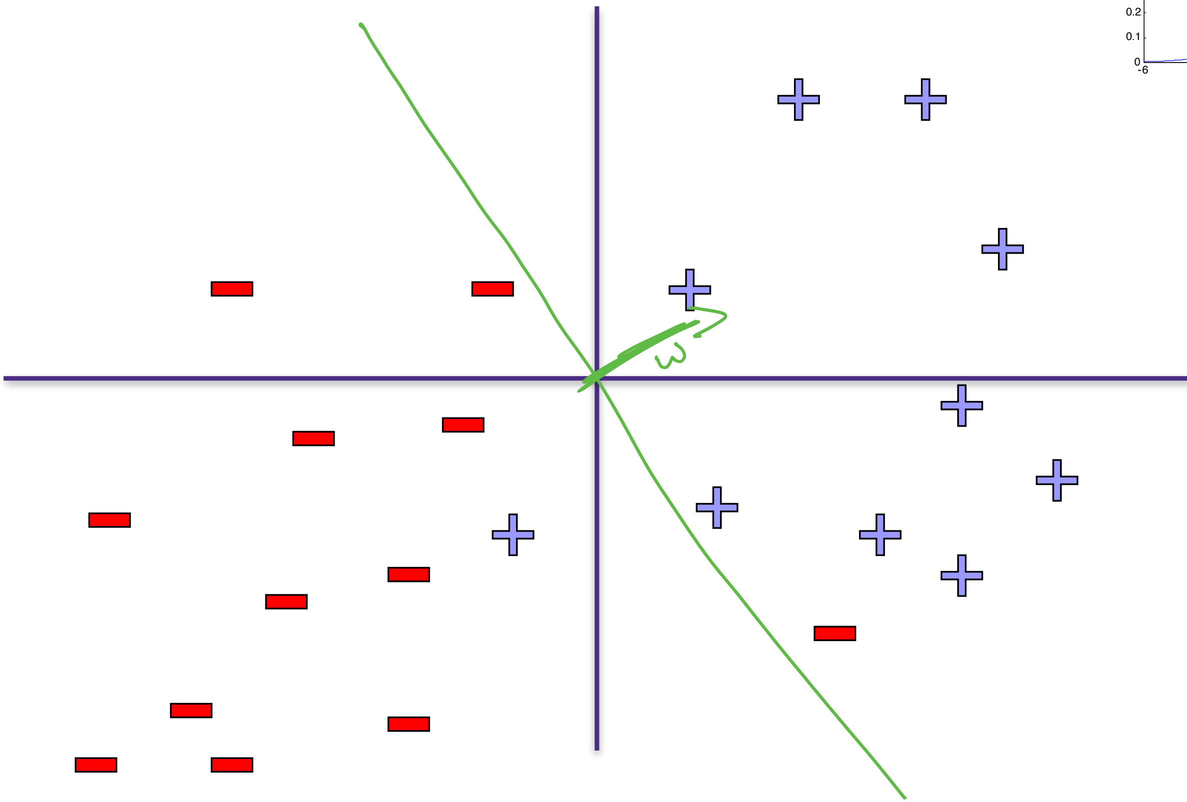
Linear Decision Rule!

$$\log\frac{\mathbb{P}(Y=1|w,X)}{\mathbb{P}(Y=\cancel{0}_{-1}|w,X)} = w_0+\sum_k w_k X_k \geq 0$$

Logistic Regression – a Linear classifier

$$\frac{1}{1+exp(-z)}$$





$$\ln \frac{P(Y=0|X)}{P(Y=1|X)} = w_0 + \sum_i w_i X_i$$

Process

Decide on a **model** make assumption

Find the function which fits the data best

Choose a loss function

Pick the function which minimizes loss on data

Use function to make prediction on new examples

Loss function: Conditional Likelihood

- **Have a bunch of iid data:** $\{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^d, \quad y_i \in \{-1, 1\}$

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y\, w^T x)}$$

$$\widehat{w}_{MLE} = \arg\max_w \prod_{i=1}^n P(y_i|x_i, w)$$

$$\boxed{= \arg\min_w \sum_{i=1}^n \log(1 + \exp(-y_i\, x_i^T w))} = \sum_{i=1}^n \ell_i(x_i, y_i, w)$$

Logistic Loss: $\ell_i(w) = \log(1 + \exp(-y_i\, x_i^T w))$

Squared error Loss: $\ell_i(w) = (y_i - x_i^T w)^2$

(MLE for Gaussian noise)

Process

Decide on a **model**

Find the function which fits the data best

Choose a loss function

Pick the function which minimizes loss on data

Use function to make prediction on new examples

Loss function: Conditional Likelihood

- **Have a bunch of iid data:** $\{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^d, \quad y_i \in \{-1, 1\}$

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y\, w^T x)}$$

$$\widehat{w}_{MLE} = \arg\max_w \prod_{i=1}^n P(y_i|x_i, w)$$

$$= \arg\min_w \sum_{i=1}^n \log(1 + \exp(-y_i\, x_i^T w)) = J(w)$$

What does $J(w)$ look like? Is it convex?

Loss function: Conditional Likelihood

- **Have a bunch of iid data:** $\{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^d, \quad y_i \in \{-1, 1\}$

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y\, w^T x)}$$

$$\widehat{w}_{MLE} = \arg\max_w \prod_{i=1}^n P(y_i|x_i, w)$$

$$= \underbrace{\arg\min_w \sum_{i=1}^n \log(1 + \exp(-y_i\, x_i^T w))}\ = \overset{\arg\min_w}{J(w)}$$

Good news: $J(\mathbf{w})$ is convex function of $\mathbf{w}$, no local optima problems

Bad news: no closed-form solution to ~~maximize~~ minimize $J(\mathbf{w})$

Good news: convex functions easy to optimize

One other concern… overfitting.

- **Have a bunch of iid data:** $\{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^d, \quad y_i \in \{-1, 1\}$

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y\, w^T x)}$$

$$\begin{aligned}\widehat{w}_{MLE} &= \arg\max_w \prod_{i=1}^n P(y_i|x_i, w) \\ &= \arg\min_w \sum_{i=1}^n \log(1 + \exp(-y_i\, x_i^T w))\end{aligned}$$

Does anyone see a situation when this minimization might overfit?

Overfitting and Linear Separability

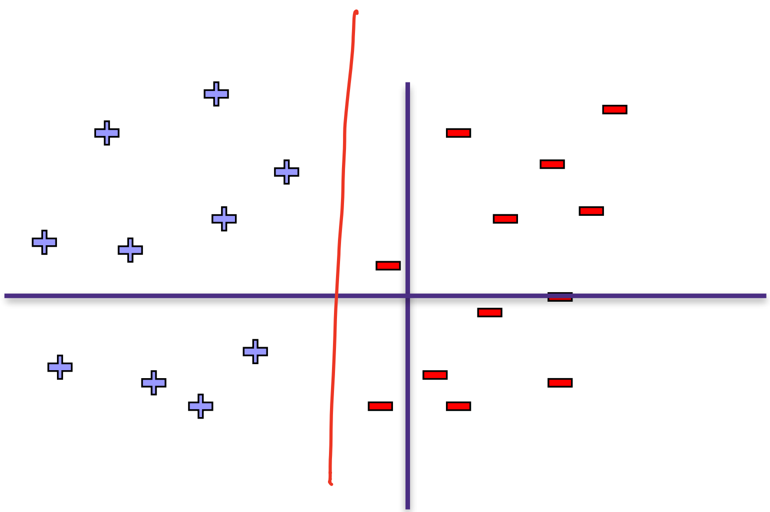
$$\arg\min_w \sum_{i=1}^{n} \log(1+\exp(-\boxed{y_i\, x_i^T\, w}))$$

always ≥ 0

Same sign

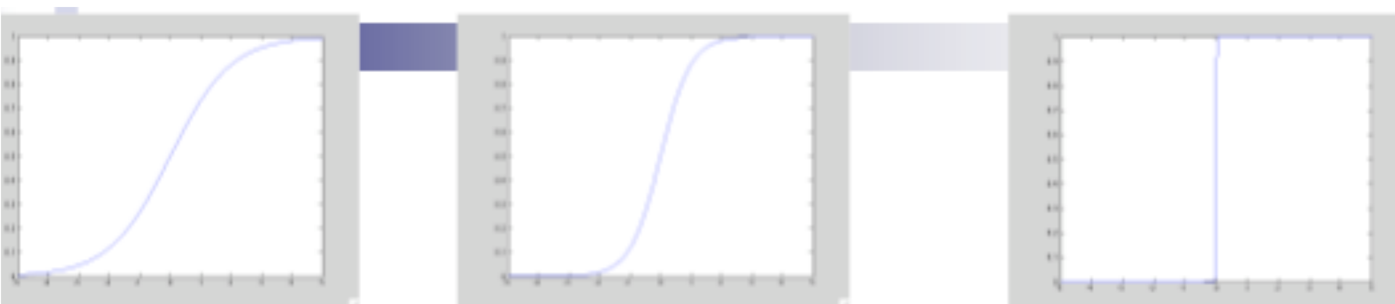
When is this loss small?

log(1+exp(− pos#))



Large parameters → Overfitting

When data is linearly separable, weights $\Rightarrow \infty$



$$\frac{1}{1+e^{-x}} \quad f(\omega) \qquad\qquad \frac{1}{1+e^{-2x}} \qquad\qquad \frac{1}{1+e^{-100x}}$$

Overfitting

Penalize high weights to prevent overfitting?

Regularized Conditional Log Likelihood

Add a penalty to avoid high weights/overfitting?:

$$\arg\min_{w,b} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i\left(x_i^T w + \underbrace{b}_{w_0})\right)\right) + \lambda||w||_2^2$$

Be sure to not regularize the offset $\cancel{b}$ w_0!

Gradient Descent

Some unfinished business…

LASSO and Logistic regression didn't have closed-form model descriptions

… we waved our hands and said "the loss functions are convex, optimize"

what did we mean by that, and how do we "optimize" a convex function?

Standard Machine Learning Problem Setup

- **Have a bunch of iid data:**

$$\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d \qquad y_i \in \mathbb{R}$$

- **Want to learn a model's parameters:**

Each $\ell_i(w)$ is convex. $\quad \arg\min_w \sum_{i=1}^n \underline{\ell_i(w)}$

the sum of convex fns is convex!

Convexity

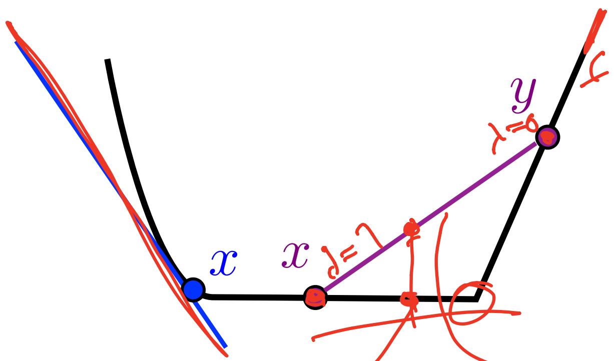
- **Have a bunch of iid data:**

$$\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d \qquad y_i \in \mathbb{R}$$

- **Want to learn a model's parameters:**

$$\sum_{i=1}^n \ell_i(w)$$

Each $\ell_i(w)$ is convex.



g is a subgradient at x if

$$f(y) \geq f(x) + g^T(y-x)$$

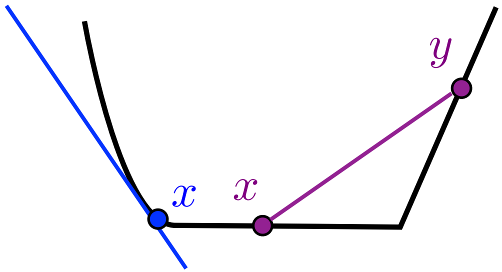
f convex:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \qquad \forall x, y, \lambda \in [0,1]$$

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) \qquad \forall x, y$$

$f(y) - f(x) \geq \nabla f(x)(y-x)$

Convexity: two equivalent definitions



g is a subgradient at x if

$$f(y) \geq f(x) + g^T(y - x)$$

f convex:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \qquad \forall x, y, \lambda \in [0,1]$$

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \qquad \forall x, y$$

$$\boxed{f(\lambda x + (1-\lambda)y)} \leq \lambda f(x) + (1-\lambda) f(y)$$

$$= \lambda f(x) + f(y) - \lambda f(y)$$

$$= \lambda (f(x) - f(y)) + f(y)$$

$\Leftrightarrow$

$$\lambda [f(x) - f(y)] \geq f(\lambda x + (1-\lambda)y) - f(y)$$

$$\frac{f(x) - f(y)}{x - y} \geq \frac{f(\lambda x + (1-\lambda)y) - f(y)}{(x-y)\lambda}, \quad \text{take} \lim_{\lambda \to 0} \frac{f(x)-f(y)}{x-y} \geq \nabla f(y)$$

(and swap x for y, thy)

Two convex loss functions

- **Have a bunch of iid data:**

$$\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d \qquad y_i \in \mathbb{R}$$

- **Want to learn a model's parameters:**

Each $\ell_i(w)$ is convex. $\qquad \sum_{i=1}^n \ell_i(w)$

Logistic Loss: $\ell_i(w) = \boxed{\log(1 + \exp(-y_i\, x_i^T w))}$

Squared error Loss: $\ell_i(w) = (y_i - x_i^T w)^2$

Least squares

- **Have a bunch of iid data:**

$$\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d \qquad y_i \in \mathbb{R}$$

- **Want to learn a model's parameters:**

$$\text{Each } \ell_i(w) \text{ is convex.} \qquad \sum_{i=1}^n \ell_i(w)$$

$$\text{Squared error Loss: } \ell_i(w) = (y_i - x_i^T w)^2$$

How does software solve: $\frac{1}{2}||\mathrm{X}w - \mathrm{y}||_2^2$

Least squares

- **Have a bunch of iid data:**

$$\{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^d \quad y_i \in \mathbb{R}$$

- **Want to learn a model's parameters:**

$$\text{Each } \ell_i(w) \text{ is convex.} \qquad \sum_{i=1}^n \ell_i(w)$$

$$\text{Squared error Loss: } \ell_i(w) = (y_i - x_i^T w)^2$$

How does software solve: $\frac{1}{2}||\mathrm{X}w - \mathrm{y}||_2^2$

…its complicated:

(LAPACK, BLAS, MKL…)

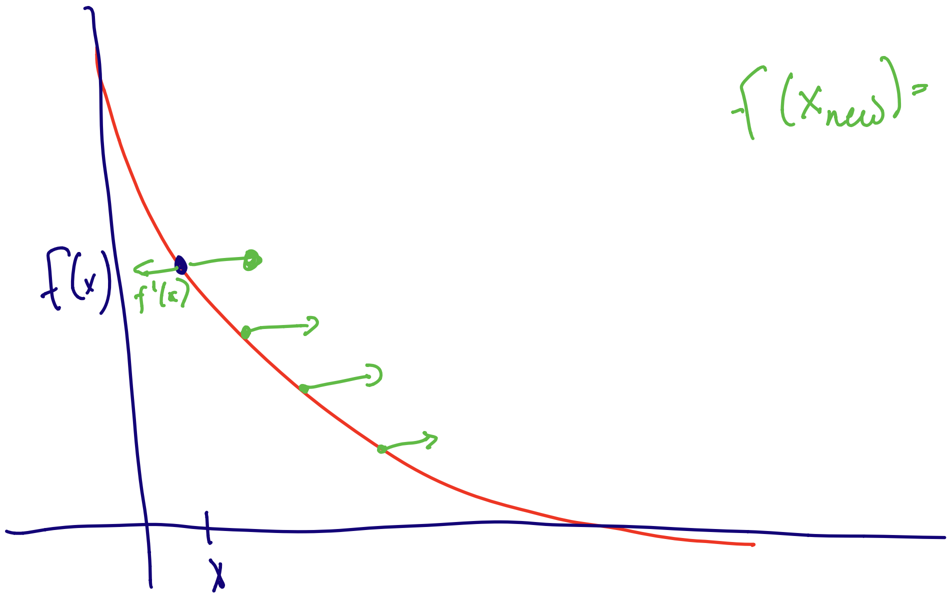
Do you need high precision?
Is X column/row sparse?
Is $\widehat{w}_{LS}$ sparse?
Is $\mathrm{X}^T\mathrm{X}$ "well-conditioned"?
Can $\mathrm{X}^T\mathrm{X}$ fit in cache/memory?

Taylor Series Approximation, 1-d

$$f(x+\delta) = f(x) + f'(x)\delta + \frac{1}{2} f''(x)\delta^2 + \ldots$$

- **Gradient descent:**

$f(x_{new}) = f_{x_{old}} - \eta \nabla f(x_{old})$



Taylor Series Approximation, d dimensions

$$f(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v + \ldots$$

- **Gradient descent:**

$$\boxed{x_{new} = x_{old} - \eta \nabla f(x_{old})}$$

$x_0 = 0$, $t = 0$

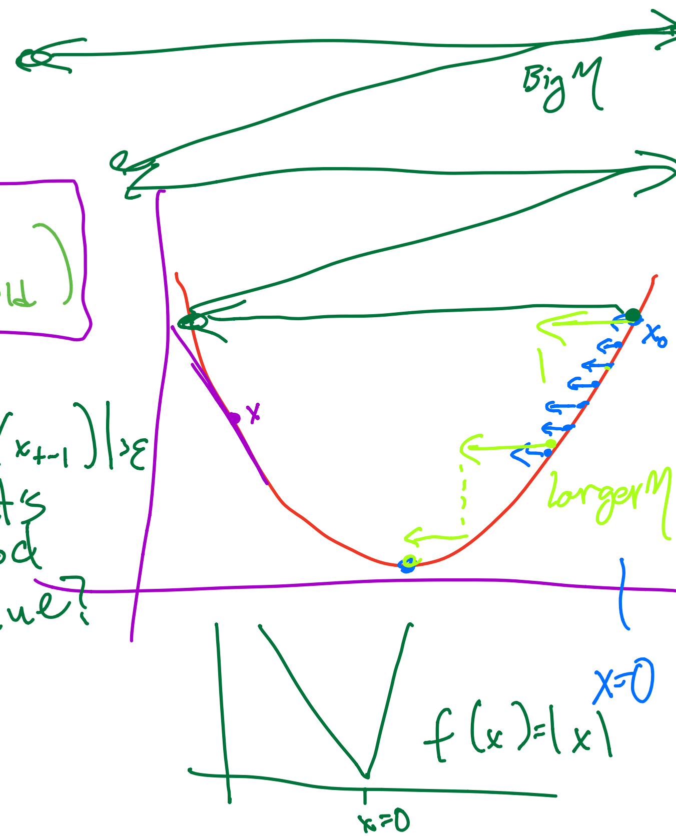
while $\|\nabla f(x_t)\|_2^2 \geq \varepsilon$

$x_{t+1} = x_t - \eta \nabla f(x_t)$

$t = t+1$

$|f(x_t) - f(x_{t-1})| > \varepsilon$

what's a good value?



Gradient Descent, LS

$$f(w) = \frac{1}{2}||\mathrm{X}w - \mathrm{y}||_2^2$$

$$\nabla f(w) = \mathbf{X}^T(\mathbf{X}w - \mathbf{y}) = \mathbf{X^T}\mathbf{X}w - \mathbf{X}^T\mathbf{y}$$

$$w_{t+1} = w_t - \eta \nabla f(w_t)$$

$$= (I - \eta \mathbf{X}^T\mathbf{X})w_t + \eta \mathbf{X}^T\mathbf{y}$$

~~If, in round t, we ended up at w_*:~~

$$\boxed{w_* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}$$

$$(w_{t+1} - w_*) = \boxed{(I - \eta \mathbf{X}^T\mathbf{X})(w_t - w_*)} \overbrace{- \eta \mathbf{X}^T\mathbf{X}w_* + \eta \mathbf{X}^T\mathbf{y}}^{=0}$$

Gradient Descent, LS

$$f(w) = \frac{1}{2}||\mathrm{X}w - \mathrm{y}||_2^2$$

$$w_{t+1} = w_t - \eta \underbrace{\boxed{\nabla f(w_t)}}_{}$$

direction

$$(w_{t+1} - w_*) = (I - \eta \mathrm{X}^T \mathrm{X})(w_t - w_*)$$

$$= (I - \eta \mathrm{X}^T \mathrm{X})^{t+1}(w_0 - w_*)$$

Gradient Descent for Logistic Regression

Loss function: Conditional Likelihood

$$\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d, \quad y_i \in \{-1, 1\}$$

$$\widehat{w}_{MLE} = \arg\max_w \prod_{i=1}^n P(y_i | x_i, w) \qquad P(Y = y | x, w) = \frac{1}{1 + \exp(-y\, w^T x)}$$

$$f(w) = \arg\min_w \sum_{i=1}^n \log(1 + \exp(-y_i\, x_i^T w))$$

$$\nabla f(w) =$$

$$\sum_{i=1}^n \nabla \log(1 + \exp(-y_i x_i^T w))$$

$$= \sum_{i=1}^n \frac{1}{1+\exp(-y_i x_i^T w)} \cdot \exp(-y_i x_i^T w)\, y_i x_i$$