Logistic regression

Comments

- Mini-review and feedback
- These are equivalent: $\mathbf{x}^\top \mathbf{w} = \mathbf{w}^\top \mathbf{x}$
- Clarification: this course is about getting you to be able to think as a machine learning expert
  - There has to be some confusion to start
  - A bit different than other courses, where need to learn a topic x (e.g., calculus) without really needing to understand why you learn topic x
  - Here you are being trained to think about how to formulate a problem, solve the problem and evaluate the problem all at once
  - Keeping it all straight takes some time

Exercise: what is c(w)?

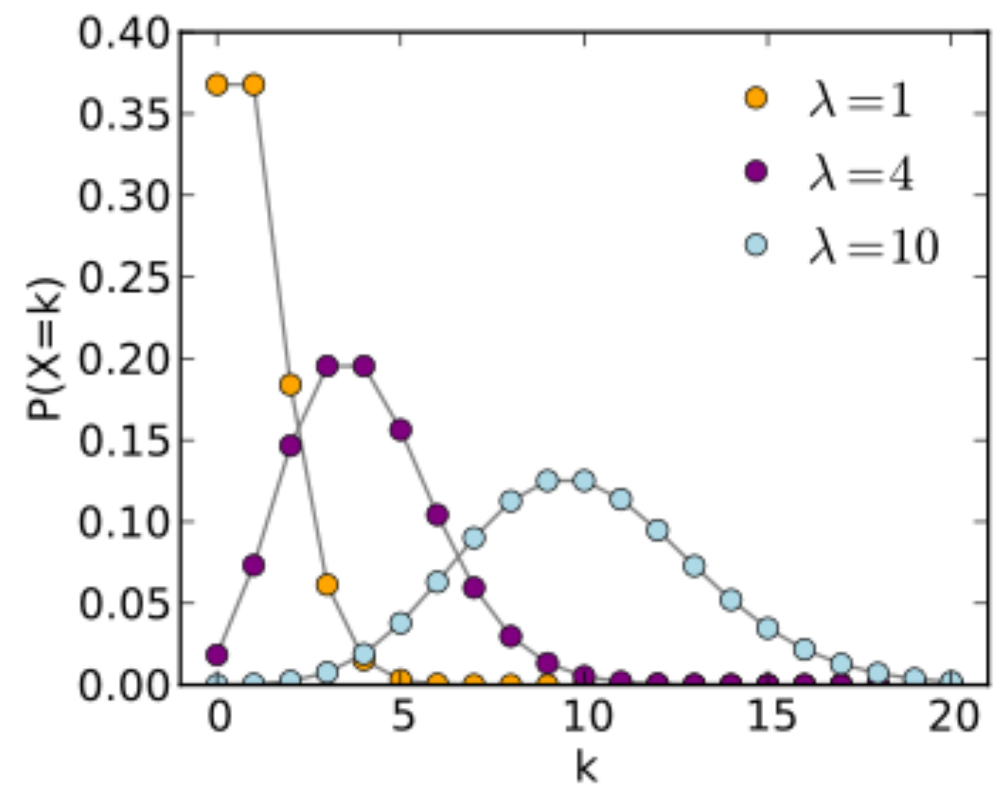
- Recall when we did MLE (maximum likelihood) for Poisson
- Assumed p(x) was Poisson, learned lambda

$$\lambda_{\mathrm{ML}} = \arg\max_{\lambda \in (0,\infty)} \{p(\mathcal{D}|\lambda)\}$$

$$\min_{w} c(w)$$

$$w = \lambda$$

$$c(w) = -\ln p(\mathcal{D}|\lambda)$$



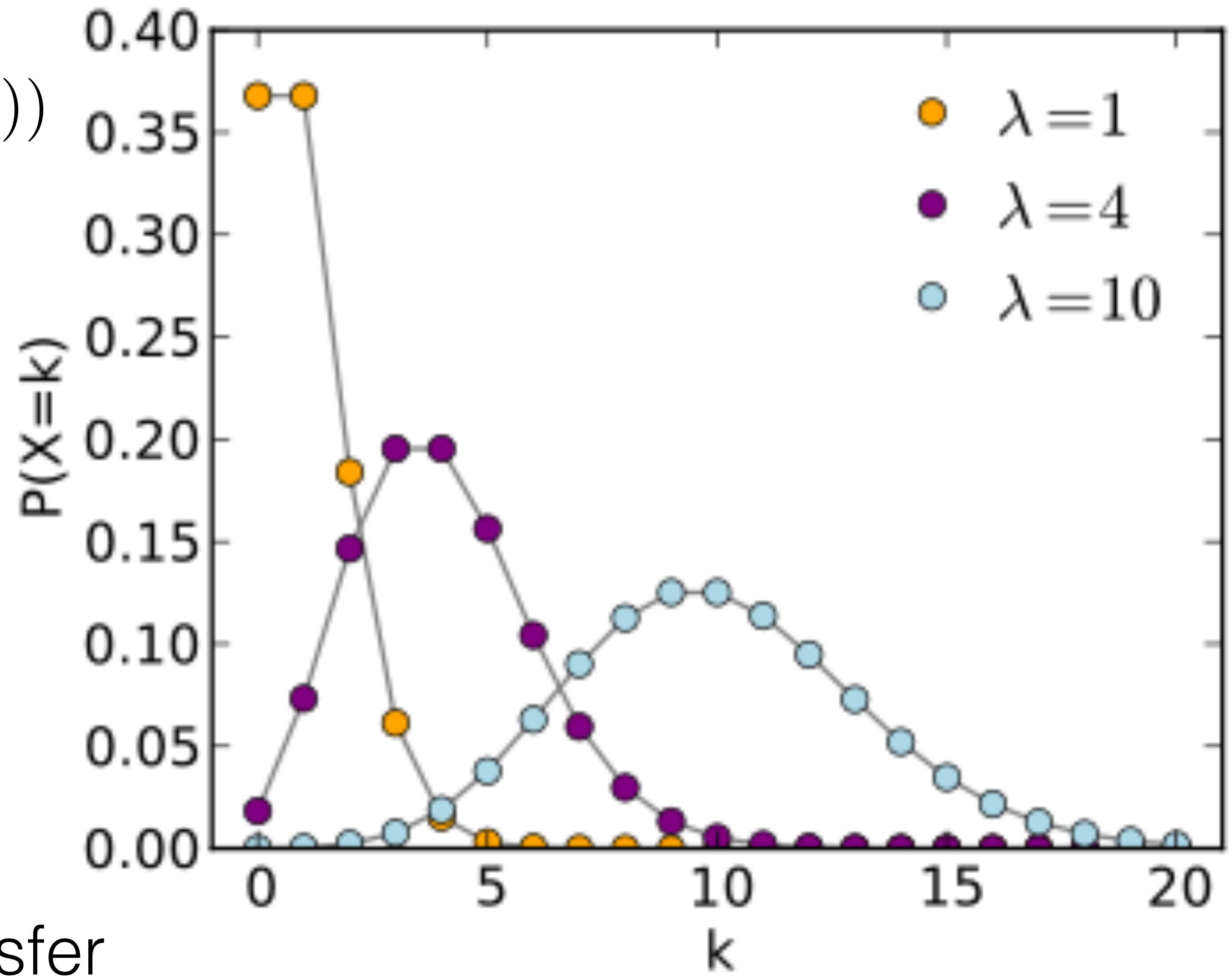
Why is Poisson regression more complicated?

- Estimated lambda for Poisson p(y), had closed form
- Estimated Poisson p(y I x), no longer have closed form!
  - Why not?!
- Why do we focus on Poisson distribution?
  - Just a canonical example, not necessarily particularly important
- Why do variable names change?
  - Previously had lambda and now have w?
  - lambda is now a function of x, i.e., $\lambda = \exp(\mathbf{x}^\top \mathbf{w})$

Poisson regression

$$p(y|\mathbf{x}) = \text{Poisson}(y|\lambda = \exp(\mathbf{x}^\top \mathbf{w}))$$

1. $\log(E[y|\mathbf{x}]) = \boldsymbol{\omega}^T \mathbf{x}$
2. $p(y|\mathbf{x}) = \text{Poisson}(\lambda)$



For exponential families, transfer f corresponds to derivate of a

$$p(y|\theta) = \exp(\theta y - a(\theta) + b(y))$$

Examples

$$\theta = \mathbf{x}^\top \mathbf{w}$$

- Gaussian distribution

$$a(\theta) = \frac{1}{2}\theta^2 \qquad f(\theta) = \theta$$

- Poisson distribution

$$a(\theta) = \exp(\theta) \qquad f(\theta) = \exp(\theta)$$

- Bernoulli distribution

$$a(\theta) = \ln(1 + \exp(\theta)) \qquad f(\theta) = \frac{1}{1 + \exp(-\theta)}$$

Exercise: How do we extract the form for the exponential distribution?

$$\lambda \exp(-\lambda y)$$

$$\lambda = f(\theta)$$

$$\theta = f^{-1}(\lambda)$$

- Recall exponential family distribution

$$p(y|\theta) = \exp(\theta y - a(\theta) + b(y))$$

- How do we write the exponential distribution this way?

- What is the transfer f?

What is c(w) for GLMS?

- Still formulating an optimization problem to predict targets y given features **x**

- The variables we learn is the weight vector **w**

- What is c(**w**)? $MLE : c(\mathbf{w}) \propto -\ln p(\mathcal{D}|\mathbf{w})$

$$\propto -\sum_{i=1}^{n} \ln p(y_i|\mathbf{x}_i\mathbf{w})$$

$$\arg\min_{\mathbf{w}} c(\mathbf{w}) = \arg\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w})$$

- Can we add regularization? How?

$$\text{Add a prior, do MAP!}$$

Extra exercises

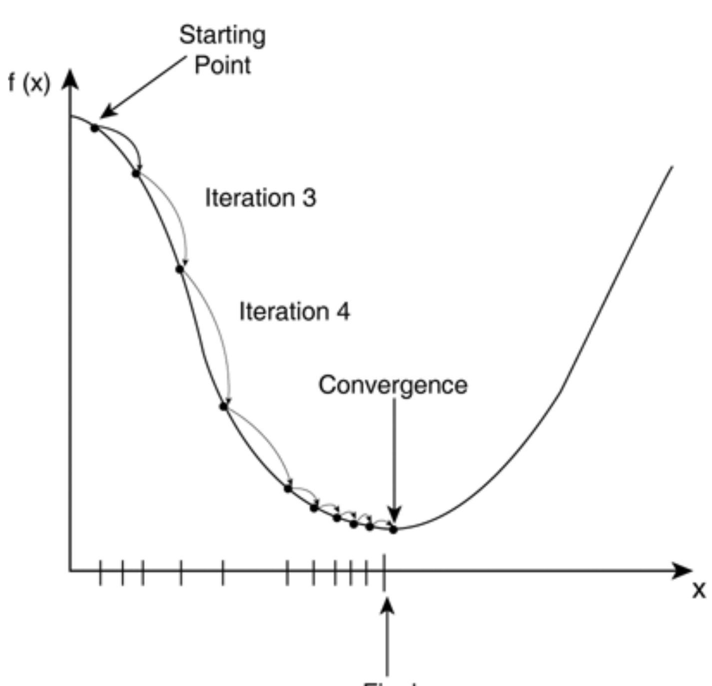
- Go through the derivation of $c(w)$ for logistic regression
- Derive Maximum Likelihood objective in Section 8.1.2

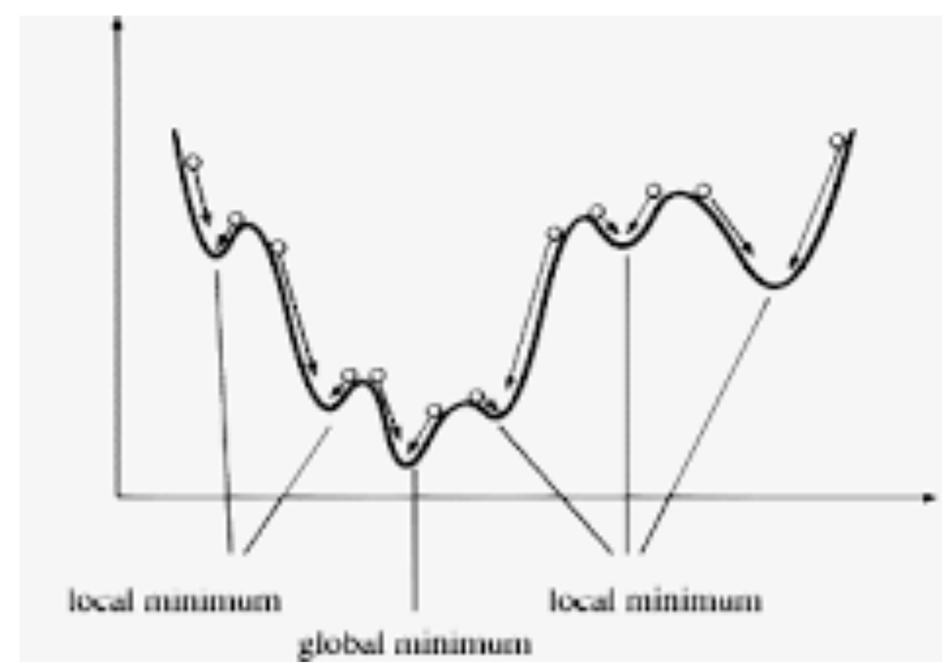
Benefits of GLMs

- Gave a generic update rule, where you only needed to know the transfer for your chosen distribution
  - e.g., linear regression with transfer f = identity
  - e.g., Poisson regression with transfer f = exp
  - e.g., logistic regression with transfer f = sigmoid
- We know the objective is convex in w!

Convexity

- Convexity of negative log likelihood of (many) exponential families
  - The negative log likelihood of many exponential families is convex, which is an important advantage of the maximum likelihood approach
- Why is convexity important?
  - e.g., why is (sigmoid(xw) - y)^2 not a good choice for binary classification?
  - Euclidean loss (squared loss) for sigmoid results in a non-convex function





How can we check convexity?

- Can check the definition of convexity

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

- Can check second derivative for scalar parameters (e.g. λ) and Hessian for multidimensional parameters (e.g., $\mathbf{w}$)
  - e.g., for linear regression (least-squares), the Hessian is $\mathbf{H} = \mathbf{X}^\top \mathbf{X}$ and so clearly positive semi-definite
  - e.g., for Poisson regression, the Hessian of the negative log-likelihood is $\mathbf{H} = \mathbf{X}^\top \mathbf{C} \mathbf{X}$ and so clearly positive semi-definite

Logistic regression

1. $\text{logit}(E[y|\mathbf{x}]) = \boldsymbol{\omega}^T \mathbf{x}$

2. $p(y|\mathbf{x}) = \text{Bernoulli}(\alpha)$

where $\text{logit}(x) = \ln \frac{x}{1-x}$, $y \in \{0,1\}$, and $\alpha \in (0,1)$

$$\alpha = p(y = 1|\mathbf{x})$$

$$g(\mathbf{x}^\top \mathbf{w}) = \text{logit}(\mathbf{x}^\top \mathbf{w})$$

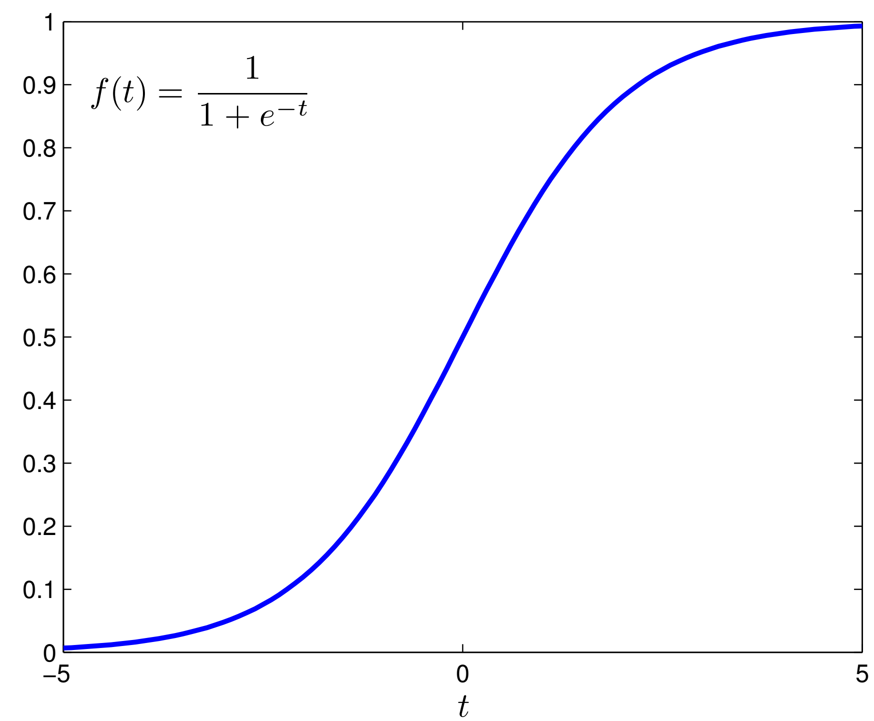
$$f(\mathbf{x}^\top \mathbf{w}) = g^{-1}(\mathbf{x}^\top \mathbf{w})$$

$$= \text{sigmoid}(\mathbf{x}^\top \mathbf{w})$$

$$= \mathbb{E}[y|\mathbf{x}]$$

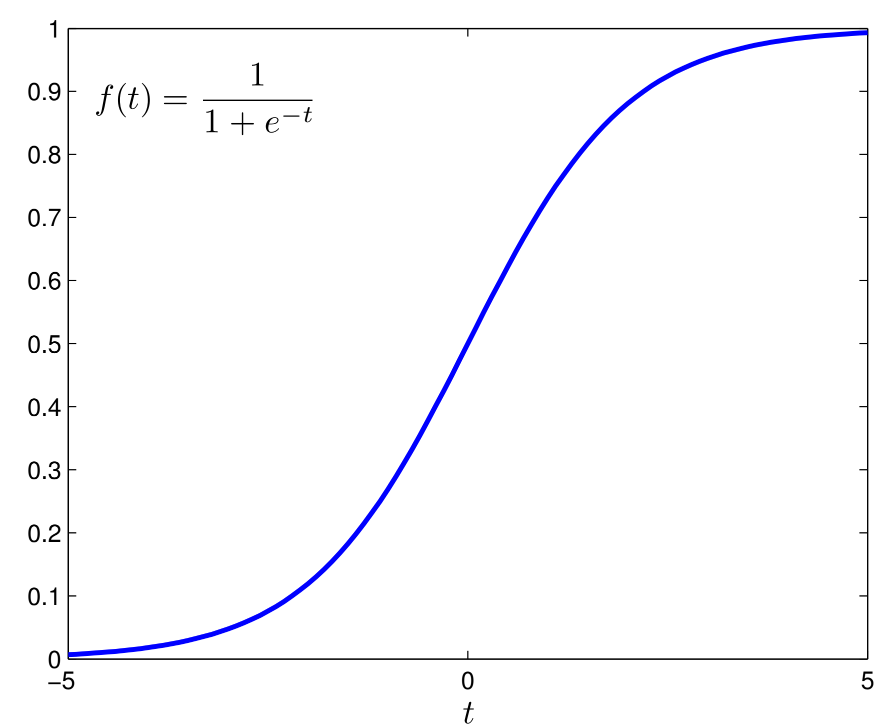
$$E[y|\mathbf{x}] = \frac{1}{1 + e^{-\boldsymbol{\omega}^T \mathbf{x}}}$$

$$p(y|\mathbf{x}) = \left(\frac{1}{1 + e^{-\omega^T \mathbf{x}}}\right)^y \left(1 - \frac{1}{1 + e^{-\omega^T \mathbf{x}}}\right)^{1-y} .$$



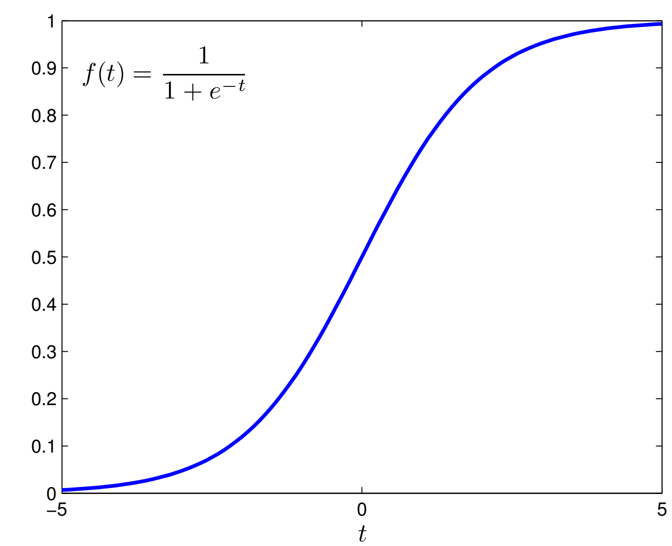
Prediction with logistic regression

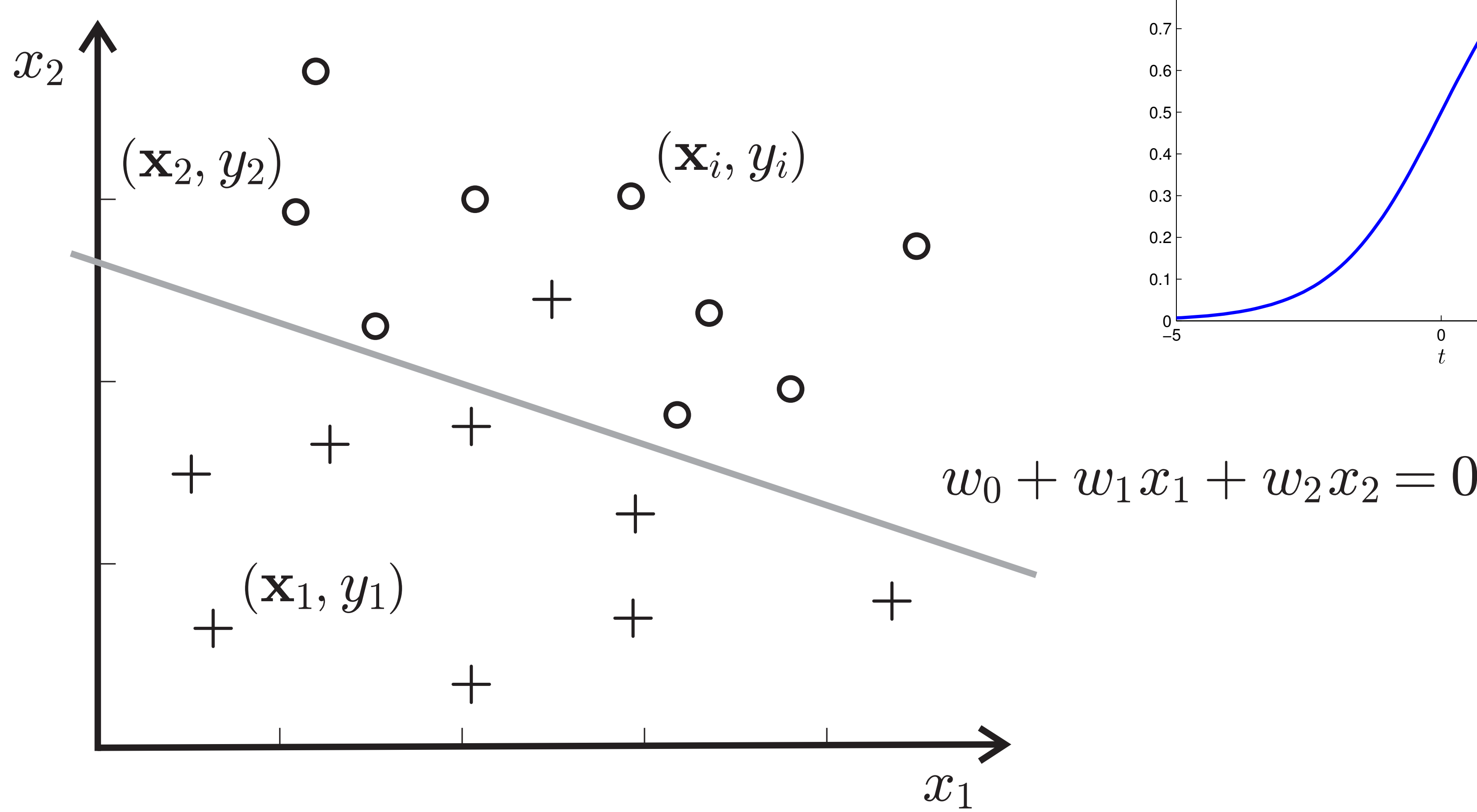
- So far, we have used the prediction f(xw)
  - eg., xw for linear regression, exp(xw) for Poisson regression
- For binary classification, want to output 0 or 1, rather than the probability value p(y = 1 | x) = sigmoid(xw)
- Sigmoid has few values xw mapped close to 0.5; most values somewhat larger than 0 are mapped close to 0 (and vice versa for 1)
- Decision threshold:
  - sigmoid(xw) < 0.5 is class 0
  - sigmoid(xw) > 0.5 is class 1



Logistic regression is a linear classifier

- Hyperplane $\mathbf{w}^\top \mathbf{x} = 0$ separates the two classes
  - P(y=1 | x, w) > 0.5 only when $\mathbf{w}^\top \mathbf{x} \geq 0.$
  - P(y=0 | x, w) > 0.5 only when P(y=1 | x, w) < 0.5, which happens when $\mathbf{w}^\top \mathbf{x} < 0$





Logistic regression versus linear regression

- Why might one be better than the other? They both use a linear approach

- Linear regression could still learn <x, w> to predict E[Y | x]

- Demo: logistic regression performs better under outliers, when the outlier is still on the correct side of the line

- Conclusion:
    - logistic regression better reflects the goals of predicting p(y=1 | x), to finding separating hyperplane
    - Linear regression assumes E[Y | x] a linear function of x!

Whiteboard

- Logistic regression
  - maximum likelihood with weightings on samples
  - optimization strategy
  - issues with minimizing Euclidean distance for sigmoid
- Multinomial logistic regression
- Next class:
  - generative approach: naive Bayes