Linear Regression

Razvan Marinescu

News:

- Quizzes seemed very popular, so we will keep them (instead of doing class exercises on Canvas)
- Quizz attendance will count towards class attendance (10%)

Linear regression – review

- In the classification tasks we've been discussing, the label space was a discrete set of classes
  - Classification, scoring, ranking, probability estimation

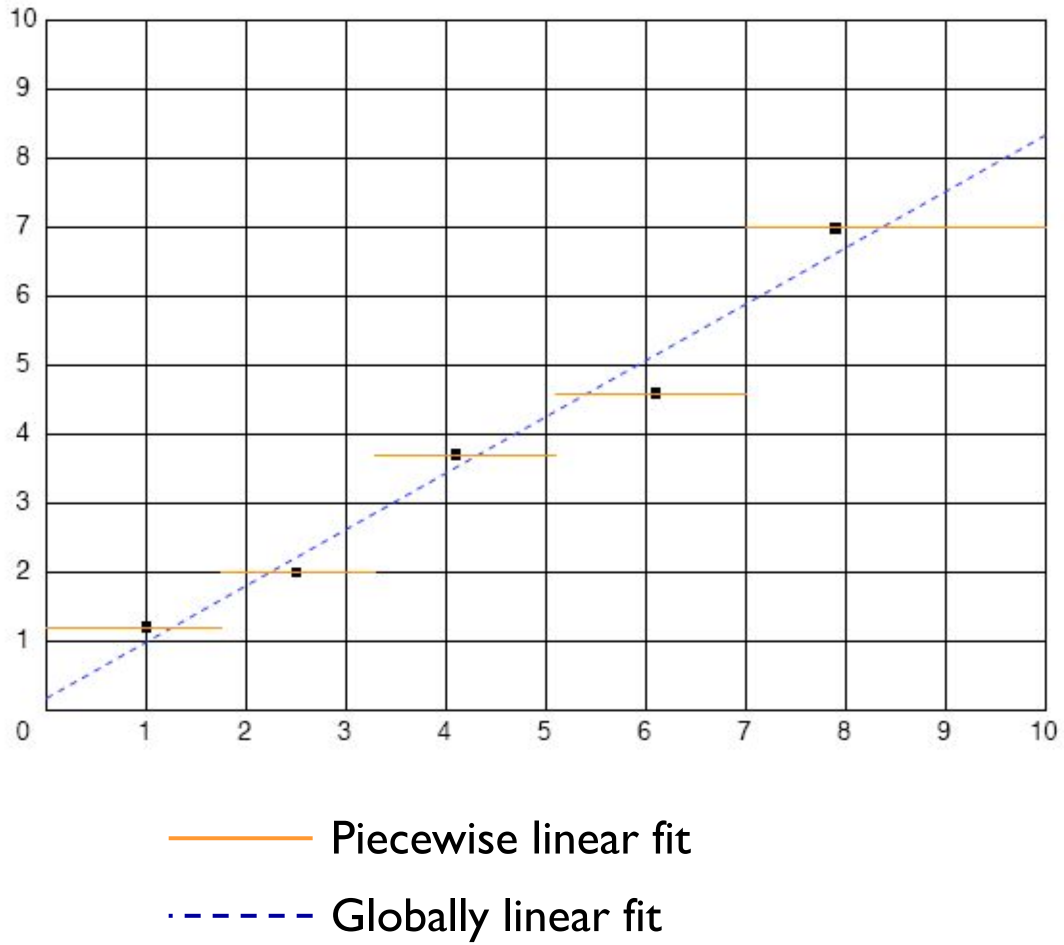
- Regression learns a function (the regressor) that is a mapping $\hat{f}$: $\mathcal{X} \rightarrow \mathbb{R}$ from examples – $f(x_i)$
  - I.e., the target variable (output) is real-valued

- Assumption: the examples will be noisy, so watch out for overfitting – we want to capture the general trend or shape of the function, not exactly match every data point

Regression example

Training data

x	$f(x)$
1.0	1.2
2.5	2.0
4.1	3.7
6.1	4.6
7.9	7.0



The regression function may or may not fit the training data exactly

Regression

What is linear regression?

Given data $\{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1..m}$

Find a θ **s.t.** $X\theta - y \approx 0$

here theta= x^-1 y.
- but x may not be invertible

Data

x_0	x_1	y
1	2	5
1	3	7
1	4	9

Data Matrix $X = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)\top} \\ \mathbf{x}^{(2)\top} \\ \mathbf{x}^{(3)\top} \end{bmatrix}$ Label $\mathbf{y} = \begin{bmatrix} 5 \\ 7 \\ 9 \end{bmatrix}$

Regression

Remark: $X\theta - y = 0$ <=> $\sum_i \left(x^{(i)}\theta - y^{(i)}\right)^2 = 0$

Regression

Remark: $X\theta - y = 0$ <=> $\sum_i \left(x^{(i)}\theta - y^{(i)}\right)^2 = 0$

Least square

argmin $J(\theta) = \frac{1}{2}\sum_{i=1}^{3}(\mathbf{x}^{(i)\top}\,\theta - y^{(i)})^2$

But

$$X\theta - \mathbf{y} = \begin{bmatrix} \mathbf{x}^{(1)} \cdot \theta \\ \vdots \\ \mathbf{x}^{(3)} \cdot \theta \end{bmatrix} - \mathbf{y} = \begin{bmatrix} \mathbf{x}^{(1)} \cdot \theta - y^{(1)} \\ \vdots \\ \mathbf{x}^{(3)} \cdot \theta - y^{(3)} \end{bmatrix}$$

Gives us the matrix form $J(\theta) = \frac{1}{2}(X\theta - \mathbf{y})^\top(X\theta - \mathbf{y})$

Regression

Least square $(X^\top X)^{-1} X^\top \mathbf{y}$

Why? $$\frac{\partial J(\theta)}{\partial \theta} = 0$$

Bias-Variance Tradeoff

Suppose that

$$y^{(i)} = f(\mathbf{x}^{(i)}) + \underbrace{\epsilon^{(i)}}_{\text{noise}}$$

where *f* is the true model

g is our model

f is true model

We're interested in a model **g**'s prediction error

$$\mathbb{E}_{\text{noise}}[(g(\mathbf{x}) - y)^2]$$

$$\mathbb{E}_{\text{noise}}[(g(\mathbf{x}) - y)^2]$$

$$= \mathbb{E}_{\text{noise}}[(g(\mathbf{x}) - f(\mathbf{x}) + f(\mathbf{x}) - y)^2]$$

recall $y = f(\mathbf{x}) + \epsilon$, so $f(\mathbf{x}) - y = -\epsilon$

$$= \mathbb{E}_{\text{noise}}[(g(\mathbf{x}) - f(\mathbf{x}) + -\epsilon)^2]$$

$$= \mathbb{E}_{\text{noise}}[(g(\mathbf{x}) - f(\mathbf{x}))^2 + \epsilon^2 - 2\epsilon(g(\mathbf{x}) - f(\mathbf{x}))]$$

$$= \mathbb{E}_{\text{noise}}[(g(\mathbf{x}) - f(\mathbf{x}))^2] + \mathbb{E}_{\text{noise}}[\epsilon^2] - 2\mathbb{E}_{\text{noise}}[\epsilon(g(\mathbf{x}) - f(\mathbf{x}))]$$

ϵ and $g(\mathbf{x}) - f(\mathbf{x})$ independent RVs so expectations multiply

$$= \mathbb{E}_{\text{noise}}[(g(\mathbf{x}) - f(\mathbf{x}))^2] + \mathbb{E}_{\text{noise}}[\epsilon^2] - 2\mathbb{E}_{\text{noise}}[\epsilon]E_{\text{noise}}[g(\mathbf{x}) - f(\mathbf{x}))]$$

$$= \mathbb{E}_{\text{noise}}[(g(\mathbf{x}) - f(\mathbf{x}))^2] + \mathbb{E}_{\text{noise}}[\epsilon^2]$$

E(e) = 0

= expected squared error to f + variance due to noise

= variance of $g(\mathbf{x})$ + (bias of $g(\mathbf{x})$)^2 + variance due to noise

$$\mathbb{E}[g(x) - f(x)]^2$$

$$= \mathbb{E}[g(x) - \bar{g} + \bar{g} - f(x))]^2$$

$$= \mathbb{E}[(g(x) - \bar{g})^2 + (\bar{g} - f(x))^2 + 2(g(x) - \bar{g})(\bar{g} - f(x))]$$

how far away is our model from true model (f(x)

= (variance of model g) + (bias of model g)^2

Last term cancels out because $\mathbb{E}[g(x) - \bar{g}] = \bar{g} - \bar{g} = 0$

Regularized Least Squares (1)

Slides from Bishop
There, they use t instead of y,
Φ(x) instead of x,
And **w** instead of \theta

- *Regularization* penalizes complexity and reduces variance (but increases bias)
- Adds a term to the squared error
- With the sum-of-squares error function and a quadratic regularizer, we get

L2 NORM OF W

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

λ is called the regularization coefficient.

- which is minimized by

$$\mathbf{w} = \left(\lambda\mathbf{I} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}.$$

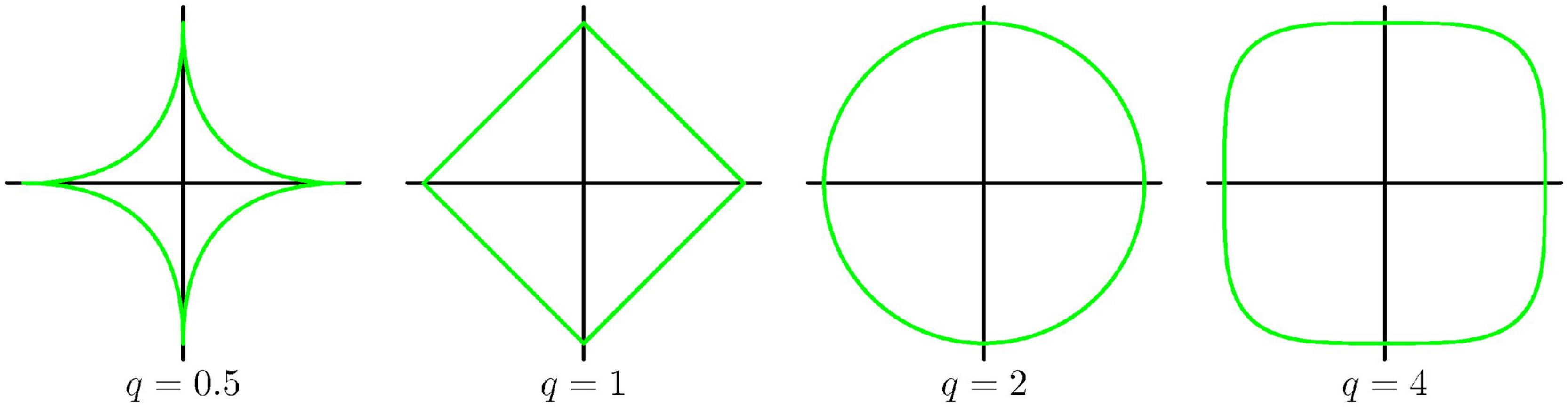
Some people view regularization as a prior on **w**’s

Regularized Least Squares (2)

With a more general regularizer, we have

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

M IS NUMBER OF VARIABLES

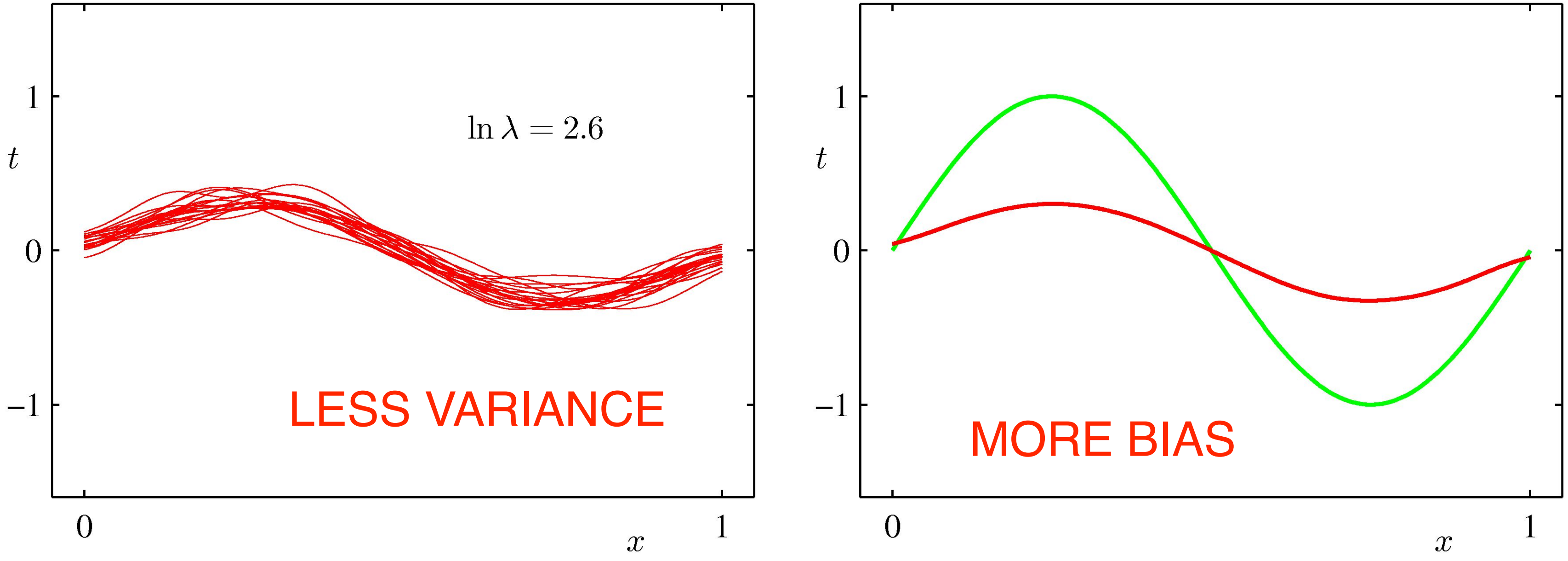


Lasso, L_1

Quadratic, L_2

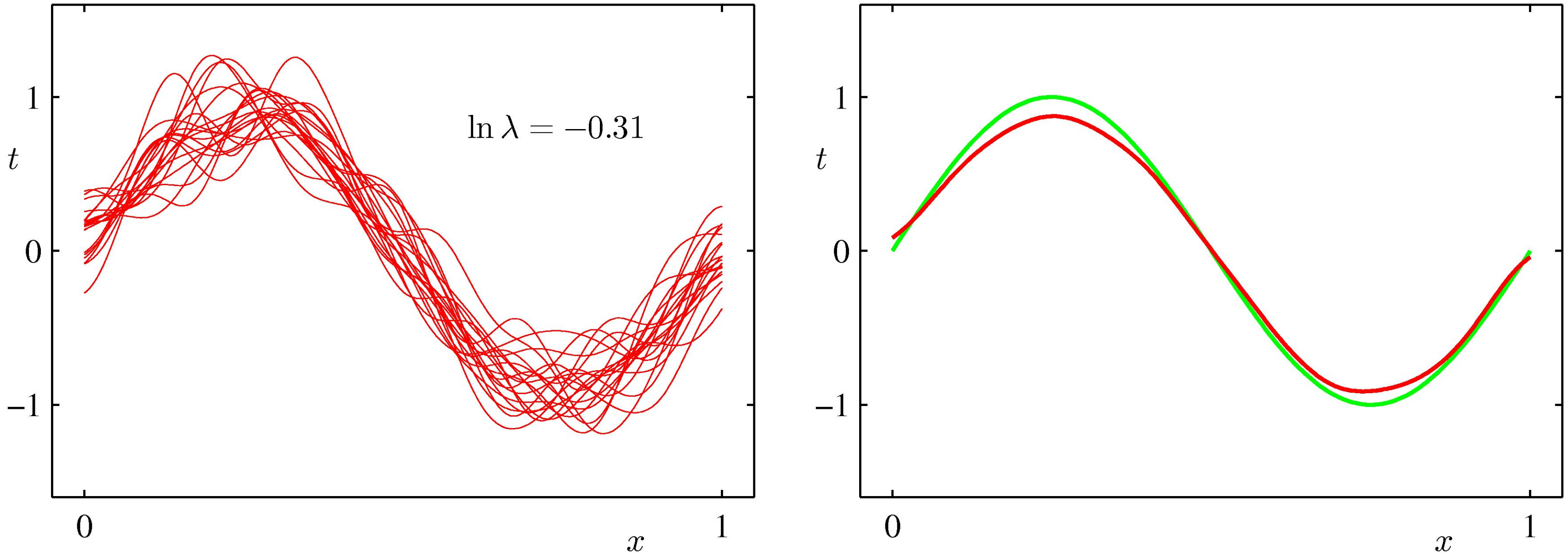
The Bias-Variance Decomposition (5)

Example: data sets from the sinusoidal, varying the degree of regularization, λ.



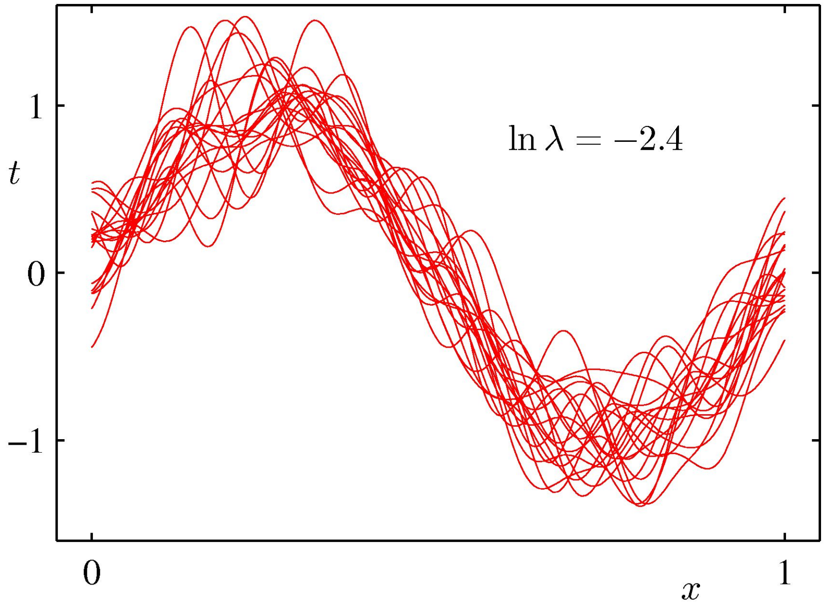
The Bias-Variance Decomposition (6)

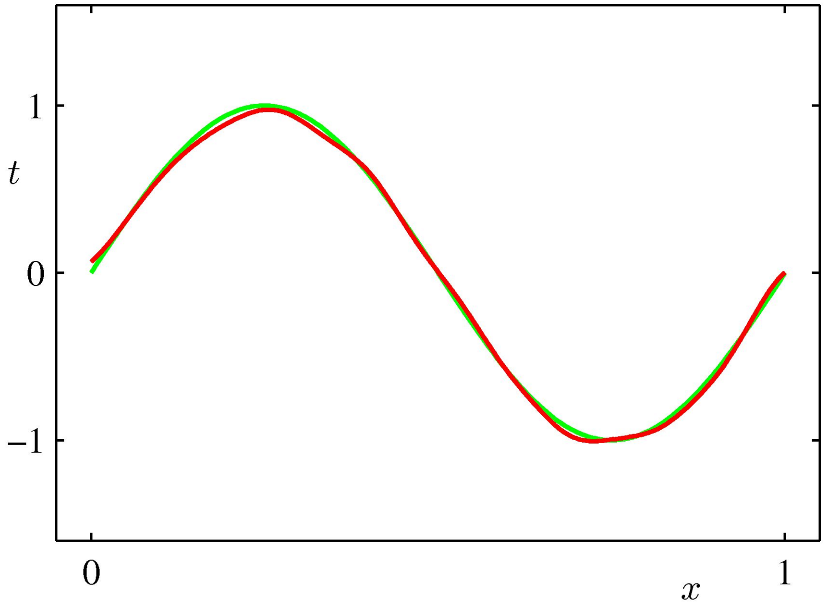
Example: 25 data sets from the sinusoidal, varying the degree of regularization, λ.



The Bias-Variance Decomposition (7)

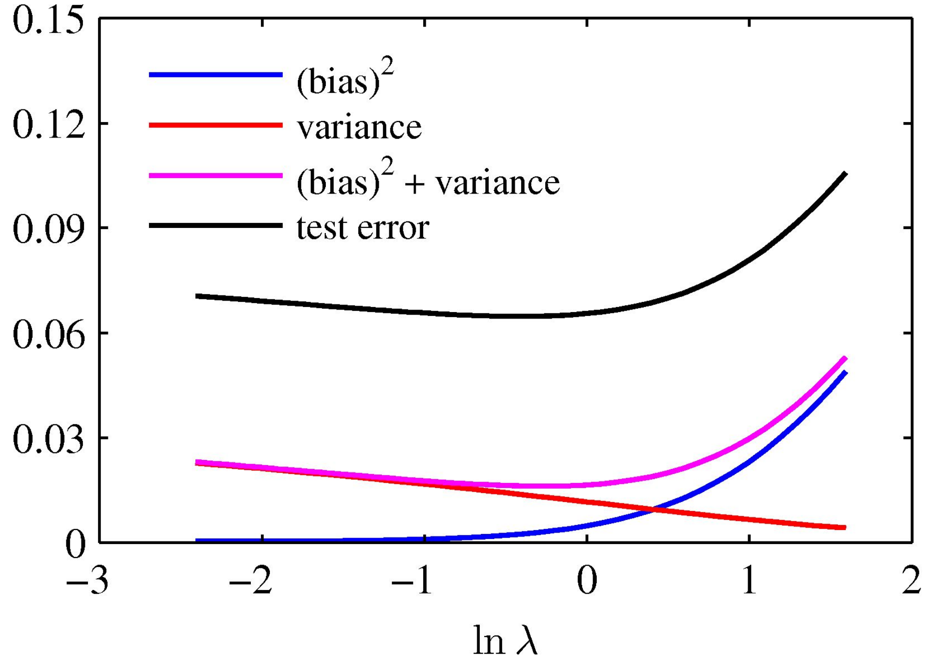
Example: 25 data sets from the sinusoidal, varying the degree of regularization, λ.





The Bias-Variance Trade-off

From these plots, we note that an over-regularized model (large λ) will have a high bias, while an under-regularized model (small λ) will have a high variance.



Logistic Regression

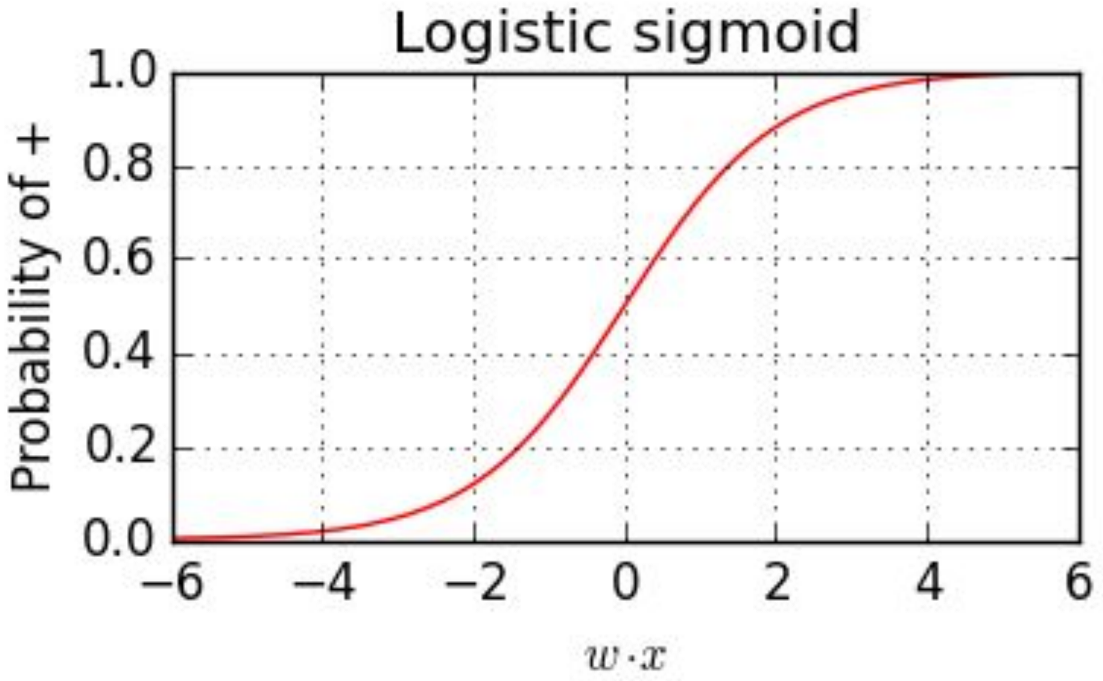
Ways to use regression to generate 2-class classifier

- Binary labels $y \in \{0, 1\}$
- Discriminative model $\mathbb{P}(y = 1|\mathbf{x}, \theta)$
- Separation hyperplane $\theta^\top \cdot \mathbf{x} = 0$
- Assume $\mathbb{P}(y = 1|\mathbf{x}, \theta) := g(\theta^\top \mathbf{x})$

Logistic Regression

What is a proper $g(\theta^\top \mathbf{x})$

- $g(-\infty) = 0$
- $g(\infty) = 1$
- $g(0) = 1/2$
- confidence of label increases as move away from boundary, so $g(\theta^\top \mathbf{x})$ is monotonically increasing
- $g(-a) = 1 - g(a)$ (symmetry, implies $g = 1/2$)



$$g(\boldsymbol{\theta} \cdot \mathbf{x}) = \frac{1}{1 + \exp(-\boldsymbol{\theta} \cdot \mathbf{x})} = \frac{\exp(\boldsymbol{\theta} \cdot \mathbf{x})}{1 + \exp(\boldsymbol{\theta} \cdot \mathbf{x})}$$

$$\mathbb{P}(y = 0 \mid \mathbf{x}; \boldsymbol{\theta}) = 1 - g(\boldsymbol{\theta}^\top \mathbf{x}) = \frac{\exp(-\boldsymbol{\theta} \cdot \mathbf{x})}{1 + \exp(-\boldsymbol{\theta} \cdot \mathbf{x})} = \frac{1}{1 + \exp(\boldsymbol{\theta} \cdot \mathbf{x})}$$

Likelihood

$$
\begin{aligned}
L(\theta) &= \mathbb{P}(y|X;\theta) \\
&= \prod_{i=1}^{m} \mathbb{P}(y^{(i)} \mid \mathbf{x}^{(i)}; \boldsymbol{\theta}) \\
&= \prod_{i=1}^{m} \underbrace{\mathbb{P}(y=1 \mid \mathbf{x}^{(i)}; \boldsymbol{\theta})^{y^{(i)}} \cdot \mathbb{P}(y=0 \mid \mathbf{x}^{(i)}; \boldsymbol{\theta})^{1-y^{(i)}}}_{\text{encodes if-test on } y} \\
&= \prod_{i=1}^{m} g(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})^{y^{(i)}} (1 - g(\boldsymbol{\theta}^\top \mathbf{x}^{(i)}))^{1-y^{(i)}}
\end{aligned}
$$

How do we maximize the likelihood?

Easier to handle the log likelihood: (a very common trick!)

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{m} y^{(i)} \log(g(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})) + (1 - y^{(i)}) \log(1 - g(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})$$

How do we solve? Gradient Descent! (next lecture)

Once you have $g(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})$

- Predict 1 if $g(\boldsymbol{\theta}^\top \mathbf{x}^{(i)}) \geq \frac{1}{2}$
- Predict 0 otherwise

JOIN QUIZ

https://quizizz.com/join

Please use your full name at sign-in.