Tuesday, Feb 17

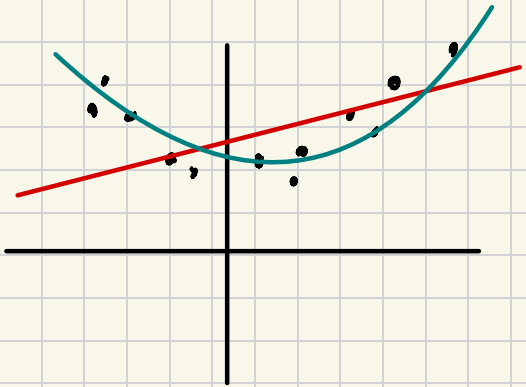
- Exam on 3/5

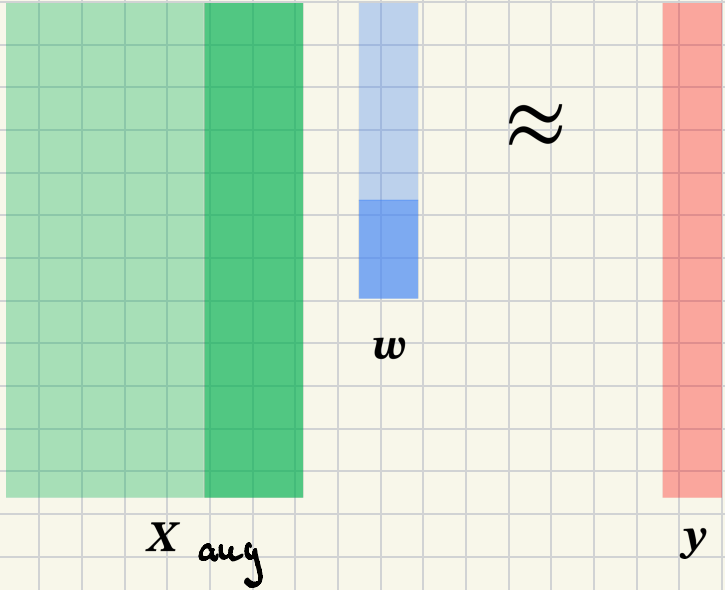
Plan

- Polynomical regression
- Logistic regression

Polynomial Regression

Can we do better than a linear fit?





$$f(x) = w_0 + w_1 x_1 + \dots + w_d x_d \qquad \text{(linear)}$$

$$f(x) = w_0 + w_1 x_1 + w_2 x_1 x_2 + w_3 x_2^2 \qquad \text{(polynomial)}$$

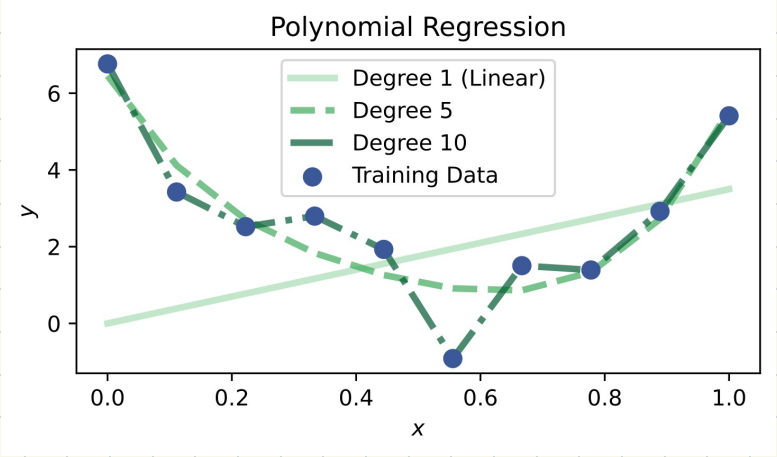
Claim:

$$\min_{w \in \mathbb{R}^{d_{aug}}} \|y - X_{aug} w\|_2^2 \leq \min_{w \in \mathbb{R}^d} \|y - Xw\|_2^2$$

Downsides of more features?

1. Time complexity
   (mitigated by kernel methods)

2. Overfitting



Generalization Error

Train → Train model

Validation → Tune hyperparams

Test → use once

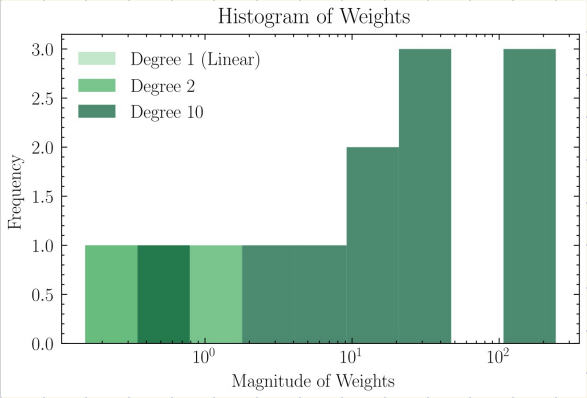
Overfit with:

1. Too many parameters
2. Too much training

Regularization

Occam's razor: Simplest explanation is usually right

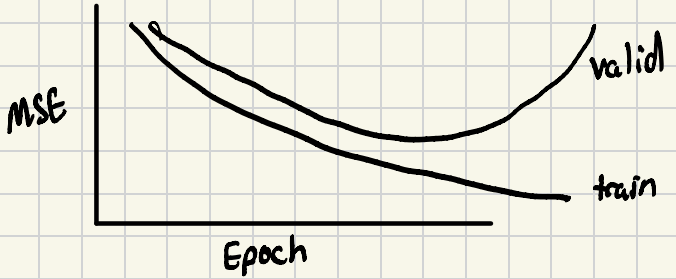
Complicated ≈ big changes
≈ big weights



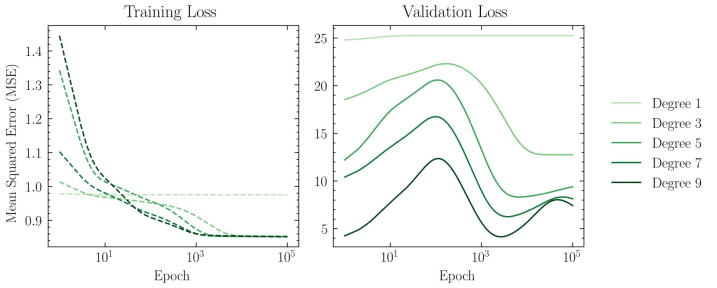
$$\underset{w}{\operatorname{argmin}} \quad \mathcal{L}(w) + \lambda \|w\|_2^2$$

Traditional View

Tradeoff between generalization and training



Double Descent



Thursday, February 19

Logistic regression!

Logistic Regression

Continuous label? Linear regression!

Discrete labels?

↳ Medicine: Sick or not

↳ Finance: Approved or not

↳ Vision: Cat or dog

$x^{(i)} \in \mathbb{R}^d$ $\quad y^{(i)} \in \{0, 1\}$

Model

Q: How do we predict binary?

Output 1 if $\langle w, x^{(i)} \rangle > 0$

Output 0 else

Q: How do we train?

Attempt #1: $|y - \langle w, x \rangle|$

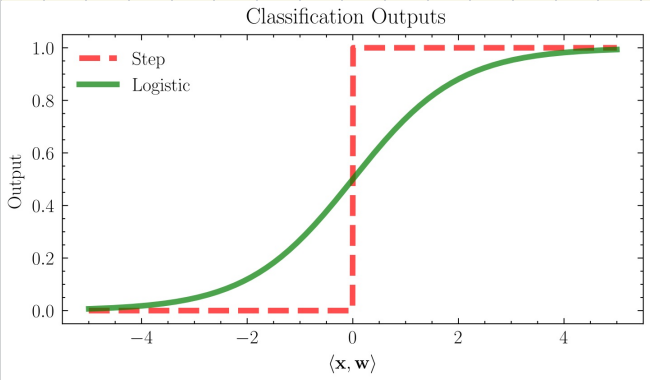
Attempt #2: $(y - \langle w, x \rangle|^2$

Attempt #3: $F(x) = \sigma(\langle w, x \rangle)$

↑ probability!

Sigmoid $\sigma : \mathbb{R} \to \mathbb{R}$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$



- smooth
- non-linear
- image is $(0,1)$

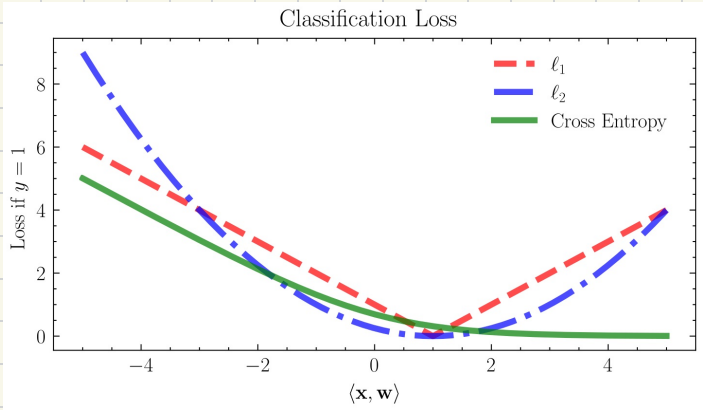
$\Rightarrow$ A probability !!

... and thresholding still works

Loss

$y \in \{0,1\}$ $\quad f(x) \in (0,1)$

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



If $y=1$, loss is $\log(f(x))$

☺ If $f(x) \approx 1$, loss is tiny

☹ If $f(x) \approx 0$, loss is huge

Maximum Likelihood Perspective

Suppose $y = \begin{cases} 1 & \text{w.p. } f(x) \\ 0 & \text{else} \end{cases}$

$$\Pr(y \mid x, w) = f(x)^y (1 - f(x))^{1-y}$$

$$\underset{w}{\operatorname{argmax}} \Pr(\text{observing data})$$

$$= \underset{w}{\operatorname{argmax}} \prod_{i=1}^{n} P(y^{(i)} \mid x^{(i)}, w)$$

$$= \ldots$$

Optimization

exact or gradient descent

Both need the gradient, so...

Lemma: $\nabla_w \mathcal{L}(w) = X^T(\sigma(Xw) - y)$

$z = \langle w, x \rangle$, $\hat{y} = \sigma(z)$

Fact 1: $\frac{\partial \sigma(z)}{\partial z} = \sigma(z)(1 - \sigma(z)) = \hat{y}(1-\hat{y})$

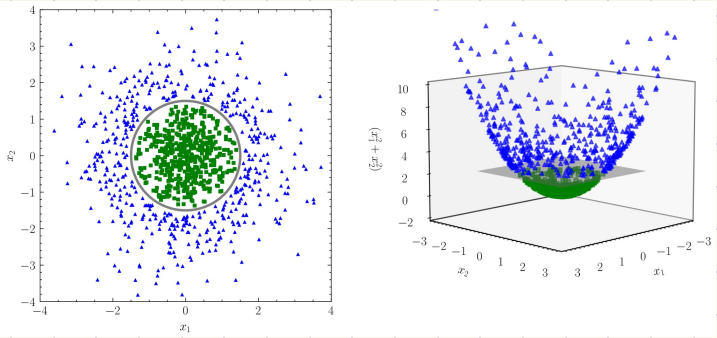
$$\ell = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$

Fact 2: $\frac{\partial \ell}{\partial \hat{y}} = -\frac{y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})} = \frac{\hat{y} - y}{\hat{y}(1-\hat{y})}$

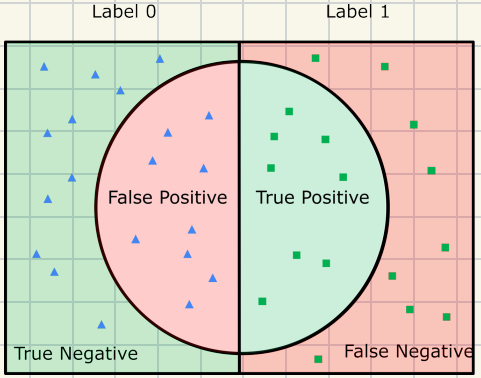
$$\nabla_w \mathcal{L}(w) = \sum_{i=1}^{n} \frac{\partial \ell^{(i)}}{\partial \hat{y}^{(i)}} \frac{\partial \hat{y}^{(i)}}{\partial z^{(i)}} \nabla_w z^{(i)}$$

$$= \dots$$

Non-linear Transformations



Measuring Error

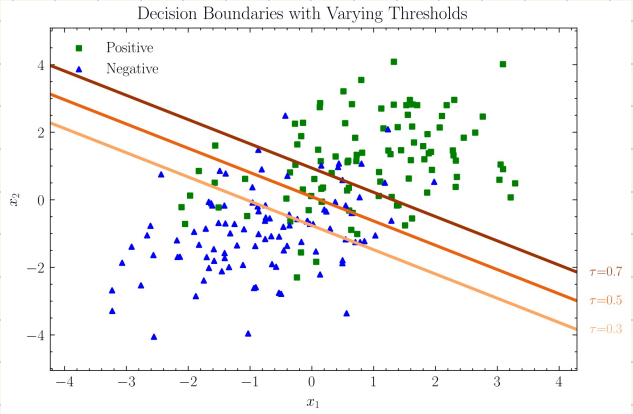


TPR = $\frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$

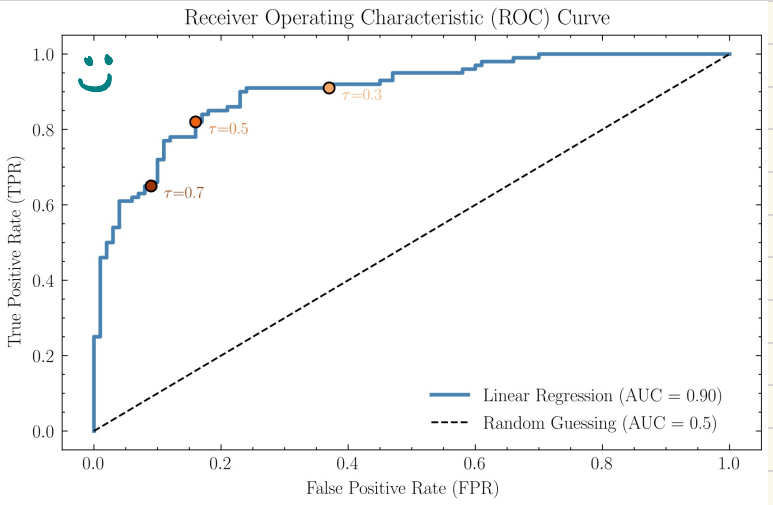
FPR = $\frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}}$

Precision = $\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$

Varying Threshold



Receiver Operating Characteristic Curve



τ

Output 1 if $\sigma(\langle w, x \rangle) > 1/2$

Output 0 else

Decreasing τ:

↑ TPR

↑ FPR