Week 6

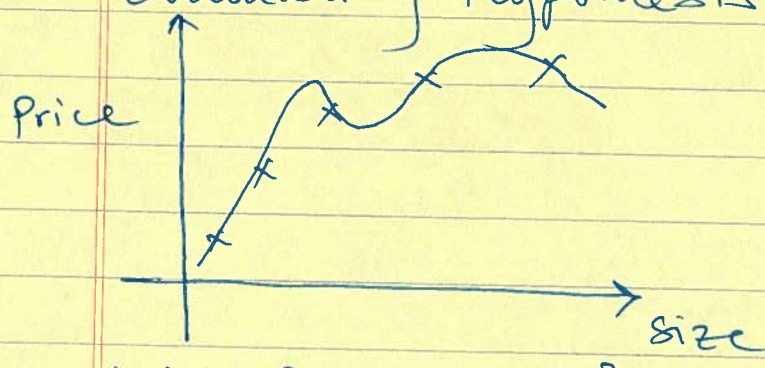
Deciding what to try next.

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{m}\theta_j^2\right]$$

- Get more training examples
- Try smaller set of features
- Try getting additional features
- Try adding polynomial features ($x_1^2, x_2^2, x_1 x_2$ ...)
- Try decreasing λ
- Try increasing λ

Machine Learning diagnostic:

Evaluating Hypothesis.



Fails to generalize to new examples

$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

70% Training set

30% Test set

Overfitting training set.
Expect training error $J(\theta)$ to be low & test error $J(\theta)$ high.

→ Learn parameter θ from traing data (minimizing training error $J(\theta)$)

- Compute test set error:

$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} \left(h_\theta(x_{test}^{(i)}) - y_{test}^{(i)} \right)^2$$

Misclassification (0/1 misclassification error)

$$err(h_\theta(x), y) = \begin{cases} 1 & \text{if } h_\theta(x) \geq 0.5, \; y = 0 \\ & \text{or if } h_\theta(x) < 0.5 \; y = 1 \end{cases} \Big\} \text{ erro} \\ \begin{cases} 0 & \text{otherwise} \end{cases}$$

$$\text{Test error} = \frac{1}{m_{test}} \sum_{i=1}^{m_{test}} err(h_\theta(x)_{test}^{(i)}, y^{(i)}).$$

d= degree of polynomial

d=1 1. $h_\theta(x) = \theta_0 + \theta_1 x$ — $\Theta^{(1)}$

d=2 2. $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 \rightarrow \Theta^{(2)}$

d=3 3. $h_\theta(x) = \theta_0 + \theta_1 x + \dots + \theta_3 x^3 \rightarrow \Theta^{(3)}$

⋮

d=3 10. $h_\theta(x) = \theta_0 + \theta_1 x + \dots + \theta_{10} x^{10} \rightarrow \Theta^{10}$

* Training set $\rightarrow$ 60%

Cross Validation set $\rightarrow$ 20% m_{cv} = # of cv examples
(cv)

Test set $\longrightarrow$ 20%. $(x^{(i)}_{cv}, y^{(i)}_{cv})$

m_{test}

Training error

$$J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

cv error

$$J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

Test Error

$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

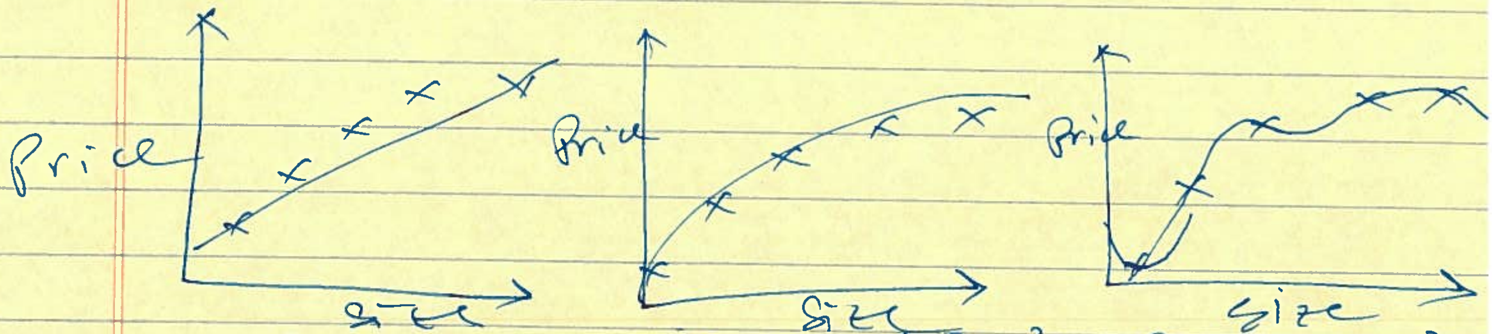
~~1. H~~

Test set to measure generalization error.

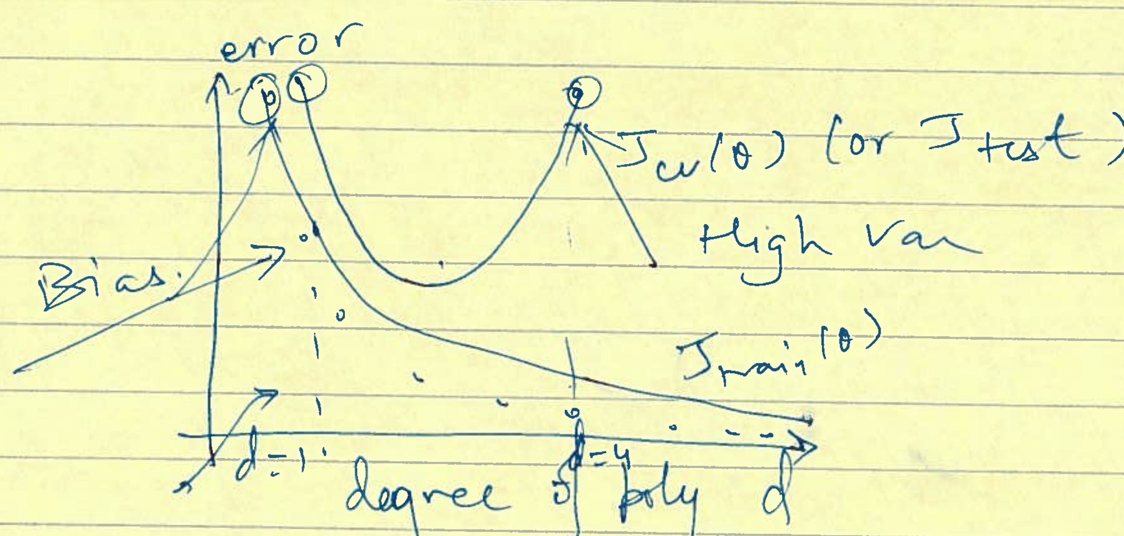
Model selection

- Traning / Validation / Test . Sets.
-

Bias Vs Variance or Both.



$\theta_0 + \theta_1 x$	$\theta_0 + \theta_1 x + \theta_2 x^2$	$\theta_0 + \theta_1 x + \theta_3 x^3 + \dots + \theta$
High Bias (Underfit)	"Just Right"	High Variance overfit
$d = 1$	$d = 2$	$d = 4$



Bias (Underfit)	Variance (Overfit)
$\left.\begin{array}{l} J_{train}(\theta) \\ J_{cv}(\theta) \end{array}\right\}$ will be high	$J_{train}(\theta)$ will be low
$J_{cv} \approx J_{train}$	$J_{cv}(\theta) \gg J_{train}(\theta)$

Regularization Bias/Variance.

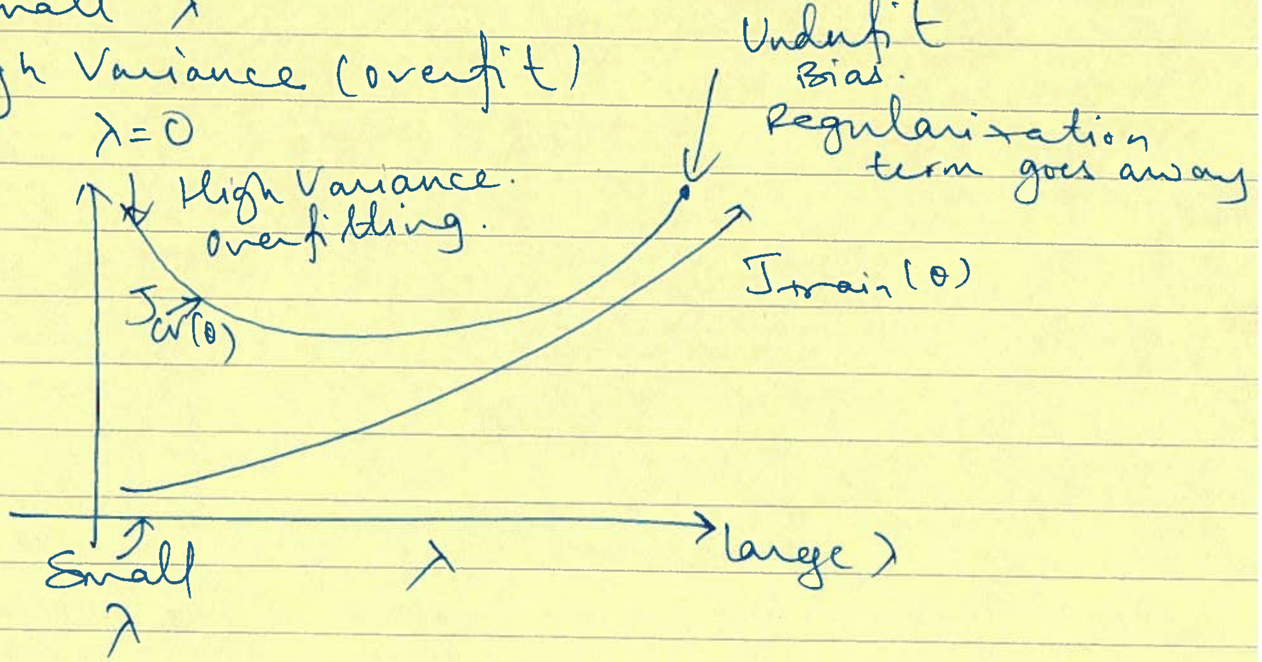
Model $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 + \frac{\lambda}{2m} \sum_{j=1}^{m} \theta_j^2$$

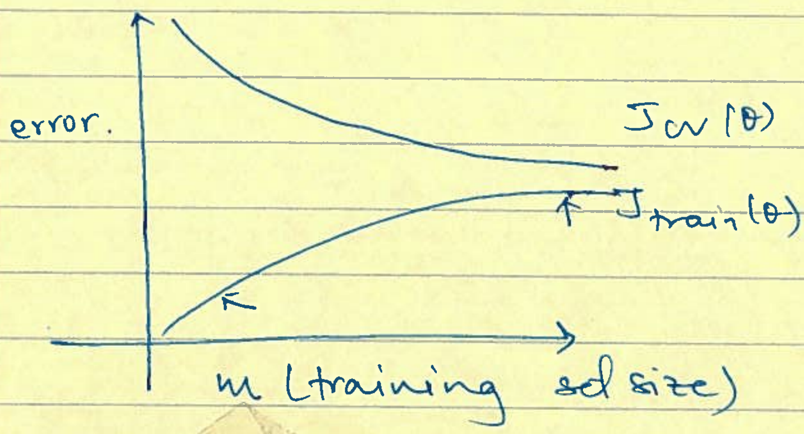
Large λ
High bias Underfit
$\lambda = 10000$, $\theta_1 \approx 0$, $\theta_2 \approx 0$
$h_\theta(x) \approx \theta_0$

Intermediate λ
"Just Right"

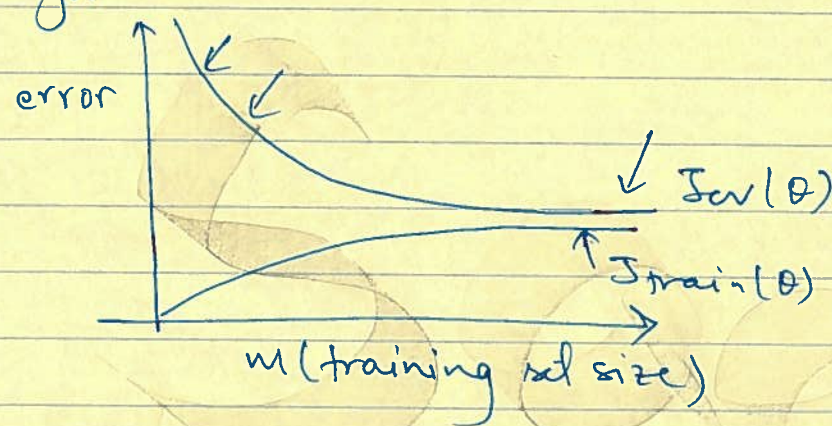
Small λ
High Variance (overfit)



Learning Curves:
Sanity check

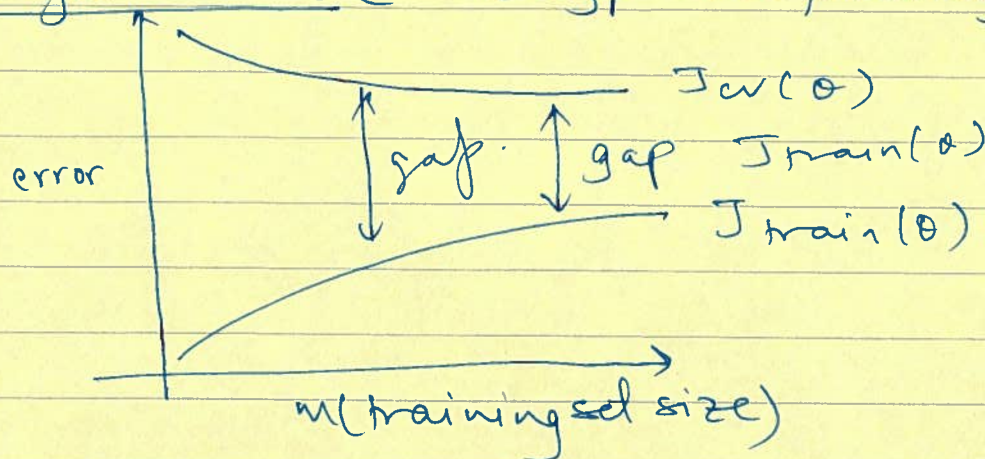


High Bias



If a learning algorithm is suffering from hig bias, getting more traning data will not (by itself) help much.

High Variance (hyp Overfitting)

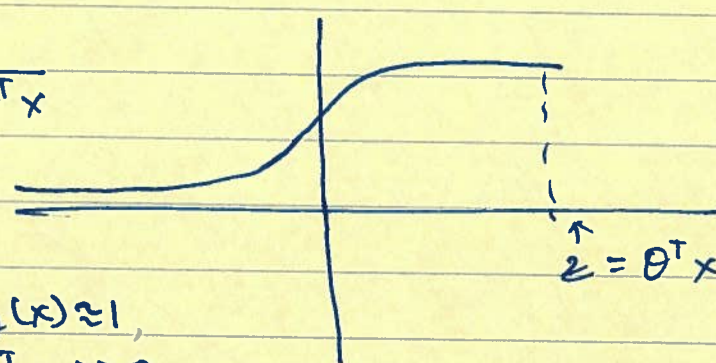


If learning algorithm is suffering from high variance, getting more training data is likely to help.

Optimization Objective

Support Vector Machine

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$



if $y = 1$, we want $h_\theta(x) \approx 1$, $\theta^T x \gg 0$

if $y = 0$ we want $h_\theta(x) \approx 0$, $\theta^T x \ll 0$

$z = \theta^T x$

Alternative view of logistic regression (x, y)

Cost of example:

$$-(y \log h_\theta(x) + (1-y) \log(1 - h_\theta(x))) \leftarrow$$

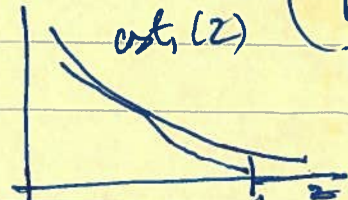
$$= \boxed{-y \log \frac{1}{1+e^{-\theta^T x}}} - (1-y) \log\left(1 - \frac{1}{1+e^{-\theta^T x}}\right)$$

If $y = 1$ (want $\theta^T x \gg 0$):

if $y = 1$ want $\theta^T x \gg 0$):

$$-\log\left(\frac{1}{1+e^{-z}}\right)$$

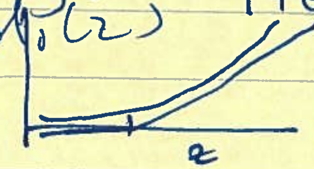
$\text{cost}_1(z)$



If $y = 0$ (want $\theta^T x \ll 0$

$$-\log\left(1 - \frac{1}{1+e^{-z}}\right)$$

$\text{cost}_0(z)$



A

Support Vector machine

$$\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)}) \text{cost}_0(\theta^T x^{(i)}) + \frac{\lambda}{2m} \sum_{j=0}^{n} \theta_j^2$$

$\min_u ((u-5)^2 + 1) \to u=5$

$\min_u 10(u-5)^2 + 10 \to u=5$

$\underline{A} + \underline{\lambda} \, \underline{B}$ if $C = \frac{1}{\lambda}$

$CA + B$

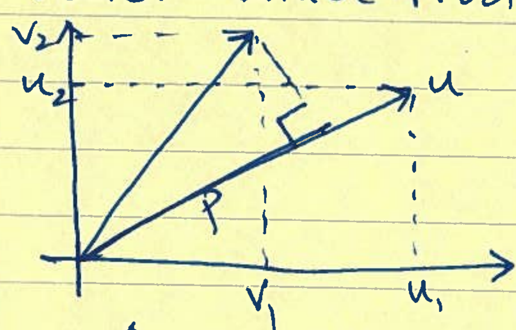
$$\min_{\theta} C \sum_{i=1}^{m} \left[y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)}) \text{cost}_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2$$

$$h_\theta(x) \begin{cases} 1 & \text{if } \theta^T x \geq 0 \\ 0 \end{cases}$$

Math behind SVM

Vector Inner Prod

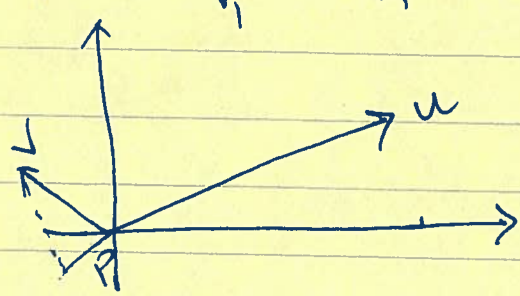
$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ $\quad v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$



$u^T v = ?$

$\|u\| = \underline{\text{length}}$ of vec u

$= \sqrt{u_1^2 + u_2^2} \in \mathbb{R}$



P = length of projection of v onto u.

$u^T v = P \cdot \|u\| = v^T u$

$= u_1 v_1 + u_2 v_2 \quad P \in \mathbb{R}$

$p < 0$

$$w = (\sqrt{w})^2$$

$$\min_\theta \frac{1}{2} \sum_{j=1}^{n} \theta_j^2 = \frac{1}{2}(\theta_1^2 + \theta_2^2) = \frac{1}{2}\left(\sqrt{\theta_1^2 + \theta_2^2}\right)^2 = \frac{1}{2}\|\theta\|^2$$

s.t. $\theta^T x^{(i)} \geq 1$ if $y^{(i)} = 1$

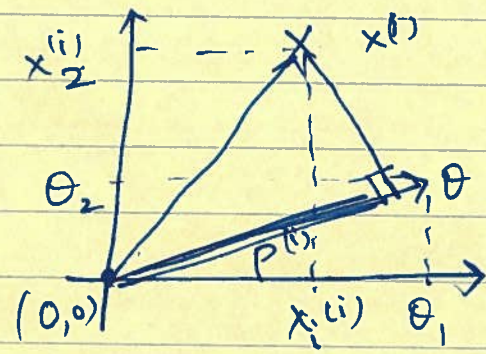
$\theta^T x^{(i)} \leq -1$ if $y^{(i)} = 0$

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

Simplication $\theta_0 = 0$ $\quad n = 2$

$\theta^T x^{(i)} = ?$

$u^T v$



$\theta^T x^{(i)} = p^{(i)} \|\theta\|$

$= \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)}$