Week 6 — Advice for applying ML

- Your dataset:
  - 60% training set
  - 20% cross validation set (CV)
  - 20% test set

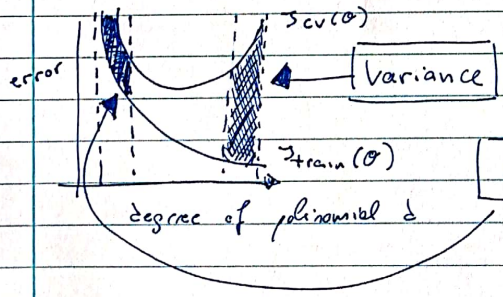
- Train / validation / test error

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

- Diagnosing bias / variance

We need to find a compromise between them.

If your algorithm is performing badly, is it a bias or variance problem?



Variance: $J_{train} \ll$, $J_{cv} \gg J_{train}$

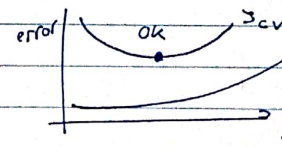
Bias: $J_{train} \gg$, $J_{cv} \approx J_{train}$

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

If $\lambda \gg \Rightarrow h_\theta(x) \approx \theta_0 \Rightarrow$ underfit

If $\lambda \gg \Rightarrow h_\theta(x) = [\dots] \Rightarrow$ overfit

How choosing the right λ? Iterating little by little through λs and minimizing $J(\theta)$ for each one. Find the minimum.



High variance
$J_{cv} >> J_{train}$

- Get more training examples
- Try smaller set of features
- Try increasing λ

High bias
$J_{cv} \approx J_{train}$
$J_{cv}, J_{train} >>$

- Try additional features
- Try adding polynomial features
- Try decreasing λ