6.036 Recitation 2/24 | Linear Regression

Agenda:

1. Linear Regression Basics
2. Measuring Error
3. Regularization
4. Solving Linear Regression
    - a. Generalized Optimization (Gradient Descent)
    - b. Closed Form Solution
    - c. Example

① Basics of Linear Regression

- Supervised Learning is essentially "learning with labeled examples"
  ↳ $\{(x^{(i)}, y^{(i)})\}$, $x^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \mathbb{R}$

  ↳ Broken up into:
  - h is a classifier • Classification: $h: \mathcal{X} \rightarrow \{C_1, C_2, \dots, C_n\}$
  - h is a generator • Structured Prediction: $h: \mathcal{X} \rightarrow \{\text{English Sentences}\}$ (for example)
  - h is a predictor • Regression: $h: \mathcal{X} \rightarrow \mathbb{R}$

- What is a predictor? (in the context of linear regression)
  ↳ A linear function of feature vectors

  i.e. $$f(x; \theta, \theta_0) = \theta \cdot x + \theta_0 = \sum_{i=1}^{d} \theta_i x_i + \theta_0$$ $\theta, x^{(i)} \in \mathbb{R}^d$

- How can we represent polynomial regression as a linear regression problem?
  – By cleverly selecting our features. Say, instead of putting just x_1 as a feature, we also included $x_1^2, x_1^3, \dots, x_1^n$. Then, we are essentially looking for the coefficients to linearly add the features $x_1, x_1^2, \dots, x_1^n$, which is the same thing as a polynomial!

- What are they useful for?
  ↳ Predicting any real value! Stock prices, test scores, salaries, house prices etc.

- What differentiates two predictors?
  ↳ θ, θ_0 contain all of the parameters

- How can we establish the best predictor? (amongst a set of them)

  $\boxed{3}$ Goals:
  - Loss function for a predictor
  - Algorithm for minimizing error
  - Good generalization to unseen data

② Measuring Error

To choose the best θ, θ_0, we need a way of scoring our model.
We can do this by analyzing <u>Empirical Risk</u>: (Assume $\theta_0 = 0$ for simplicity)

$$R_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \text{Loss}(y^{(t)} - \theta \cdot x^{(t)})$$

↑ Average Prediction Error over training set

Note that this formula contains a "Loss" function. This may be any sensible function for the particular task, but we will use the typical <u>Squared Error</u>:

| also known as Least Squares Criterion | → $\text{Loss}(z) = \frac{z^2}{2}$ |

This "$\frac{1}{2}$" term is simply for when we take a derivative so the **2** from the exponent cancels out!

- Why quadratic?
  - Symmetric (i.e. guessing low = guessing high)
  - Big errors are <u>very</u> bad relative to smaller ones.

Think back to our 3 goals. We don't "really" want to minimize Empirical Risk!
→ We want a low <u>generalization error</u>: i.e. test set error

- Where does generalization error come from?
  - <u>Estimation Error</u>: Noisy data, insufficient training datasize, overfitting...
  - <u>Structural Error</u>: Incorrect model choice (e.g. relationship is nonlinear)
    ↑ Would occur even with ∞ training examples

<u>Tradeoff</u>: More complex models are harder to train and do not generalize as well, (overfit) but can describe more complex data.

How do we combat this? Check the next page to find out!!

3) Regularization:

- Regularization is a technique used to solve overfitting in models
  - ★ General Idea (simplified): Simpler models generalize better ★
- How can we simplify our model?
  - Remove parameters (features - don't get rid of useful ones though!)
  - "Regularize Parameters" - Ridge Regression

Ridge Regression: Add an L2-Regularization term (i.e. $\lambda \frac{\|\theta\|^2}{2}$) to the error function

Again, $\boxed{1/2}$ for gradients!

★ $\boxed{\text{We wish to minimize this}}$

$$J_{n,\lambda}(\theta) = R_n(\theta) + \frac{\lambda}{2}\|\theta\|^2$$

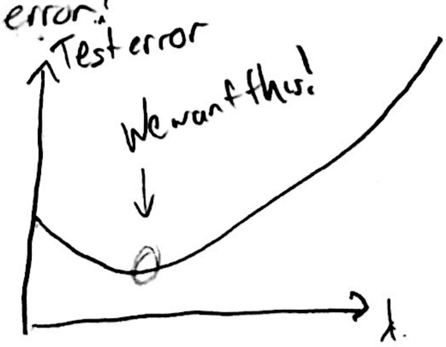
$$= \frac{1}{n}\sum_{t=1}^{n} (y^{(t)} - \theta \cdot x^{(t)})^2/2 + \frac{\lambda}{2}\|\theta\|^2$$

Comments: This will try and lower the parameters of θ as much as possible

→ Higher λ ⇒ More importance in lowering parameters (θ)

→ Lower λ ⇒ More importance in lowering training loss (empirical risk)

- Why are lower parameters better?
  - ↳ Smoother predictor, i.e. Small changes in input ⇒ Small changes in output
    (noise)
- How does λ affect training error, test error?



- Choose λ using a validation set (subset of training set used for tuning λ)

(4) Solving Linear Regression

- Remember we wish to minimize:

$$J_{n,\lambda}(\theta) = R_n(\theta) + \frac{\lambda}{2}\|\theta\|^2$$

Method $\boxed{a}$: (Stochastic) Gradient Descent

This will be hard to recompute on every update to θ! /slow

$$\theta^{(k+1)} = \theta^{(k)} - \eta_k \nabla J_{n,\lambda}(\theta)$$

$\Downarrow$

Take advantage of structure of J!

$$J_{n,\lambda}(\theta) = \frac{1}{n}\sum_{t=1}^{n} \left(y^{(t)} - \theta \cdot x^{(t)}\right)^2/2 + \frac{\lambda}{2}\|\theta\|^2$$

Since $J_{n,\lambda}$ is a sum, we can do one feature at a time!

$$\text{Regularized Loss for } x^{(t)}, y^{(t)} = \frac{\left(y^{(t)} - \theta \cdot x^{(t)}\right)^2}{2} + \frac{\lambda}{2}\|\theta\|^2$$

$$\nabla_\theta\left(\frac{\left(y^{(t)} - \theta \cdot x^{(t)}\right)^2}{2} + \frac{\lambda}{2}\|\theta\|^2\right)$$

$$= -\left(y^{(t)} - \theta \cdot x^{(t)}\right)x^{(t)} + \lambda\theta$$

$\Rightarrow$ Update Equation: $\boxed{\theta^{(k+1)} \leftarrow (1-\lambda\eta_k)\theta^{(k)} + \eta_k\left(y^{(t)} - \theta^{(k)} \cdot x^{(t)}\right)x^{(t)}}$

To turn this into an algorithm:

$\theta^{(0)} = \vec{0}$

for i in $\{1, \ldots, \text{num\_iterations}\}$:

  pick t randomly in $\{1, \ldots, n\}$:

  $\theta^{(k+1)} \leftarrow (1-\lambda\eta_k)\theta^{(k)} + \eta_k\left(y^{(t)} - \theta^{(k)} \cdot x^{(t)}\right)x^{(t)}$

4 (cont.)

b Closed Form Solution

Acceptable b/c convexity of Linear Regression

In order to minimize $J_{n,\lambda}(\theta)$, we can look for when $\nabla_\theta J_{n,\lambda}(\theta) = 0$

i.e. let's find a $\hat{\theta}$ for which $\nabla_\theta J_{n,\lambda}(\theta)_{\theta=\hat{\theta}} = 0$

$$\nabla_\theta J_{n,\lambda}(\theta)_{\theta=\hat{\theta}} = \left(\frac{1}{n}\sum_{t=1}^{n} (y^{(t)} - \theta \cdot x^{(t)})^2/2 + \frac{\lambda}{2}\|\theta\|^2\right)_{\theta=\hat{\theta}}$$

$$= \frac{1}{n}\sum_{t=1}^{n} -(y^{(t)} - \hat{\theta}\cdot x^{(t)})x^{(t)} + \lambda\hat{\theta}$$

$$= -\frac{1}{n}\sum_{t=1}^{n} y^{(t)}x^{(t)} + \frac{1}{n}\sum_{t=1}^{n}(\hat{\theta}\cdot x^{(t)})x^{(t)} + \lambda\hat{\theta}$$

$$= -\frac{1}{n}\sum_{t=1}^{n} y^{(t)}x^{(t)} + \frac{1}{n}\sum_{t=1}^{n} x^{(t)}(x^{(t)})^T\hat{\theta} + \lambda\hat{\theta}$$

$$= \underbrace{-\frac{1}{n}\sum_{t=1}^{n} y^{(t)}x^{(t)}}_{\vec{b}} + \underbrace{\left(\frac{1}{n}\sum_{t=1}^{n} x^{(t)}(x^{(t)})^T + \lambda I\right)}_{A}\hat{\theta}$$

$$\Rightarrow A\vec{\theta} - \vec{b} = \vec{0}$$

$$\Rightarrow \boxed{\vec{\theta} = A^{-1}\vec{b}}$$

☆ We assumed A was invertible. Under what conditions might A not be invertible?

- This is left to the reader as a proof will not fit in the margin :)

④ (cont.)

$\boxed{c}$ Example Polynomial Regression Problem: (Without Regularization)

Say we have the points:

$(x, y) \in \{(2,2), (3,1), (4,2) (5,5)\}$ and we wish to fit them with the quadratic of best fit.

To construct our feature vectors for such a problem they should contain a 1, and feature. Then, we will essentially be looking for the coefficients to minimize the error:

$$R_n(\theta) = \frac{1}{n} \sum_{t=1}^{4} (y^{(t)} - \theta \cdot x^{(t)})^2 / 2$$

If we take a closer look at the value of $\theta \cdot x_1^{(t)}$ then we see that it is:

$$\theta \cdot x^{(t)} = \theta_1 \cdot \overset{1}{X_1} + \theta_2 \overset{x_t}{X_2} + \theta_3 \cdot \overset{x_t^2}{X_3}$$

$$= \underbrace{\theta_3 x_t^2 + \theta_2 x_t + \theta_1}$$ which is the quadratic we were looking for (when $\theta = \theta^*$)

$\Rightarrow$ Using the closed form solution, $A = \frac{1}{n} X^T X$, $\vec{b} = \frac{1}{n} X^T \vec{y}$

$$X = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} 2 \\ 1 \\ 2 \\ 5 \end{bmatrix}$$

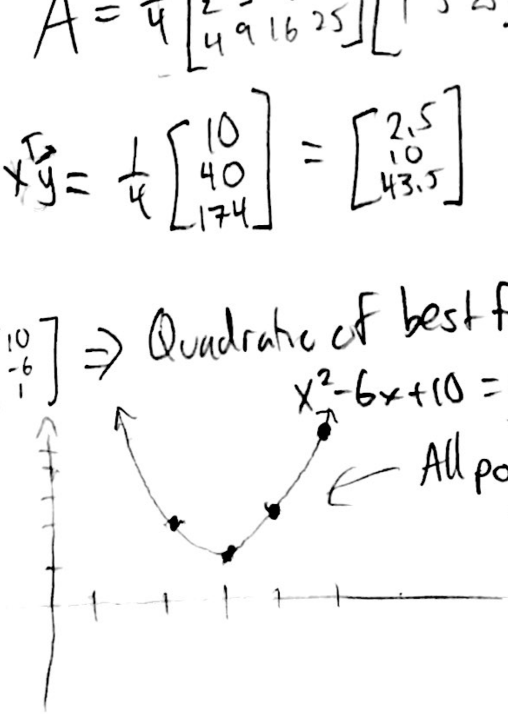
(Rows are feature vectors)

$$A = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 3 & 4 & 5 \\ 4 & 9 & 16 & 25 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix} = \begin{bmatrix} 1 & 3.5 & 13.5 \\ 3.5 & 13.5 & 56 \\ 13.5 & 56 & 244.5 \end{bmatrix}$$

$$\vec{b} = \frac{1}{n} X^T \vec{y} = \frac{1}{4} \begin{bmatrix} 10 \\ 40 \\ 174 \end{bmatrix} = \begin{bmatrix} 2.5 \\ 10 \\ 43.5 \end{bmatrix}$$

$A^{-1} \vec{b} = \begin{bmatrix} 10 \\ -6 \\ 1 \end{bmatrix} \Rightarrow$ Quadratic of best fit is!

$x^2 - 6x + 10 = y$



All points are on the curve!

Math works!