BLG 561 E Deep Learning

CRN 14002

25.09.2018

(Intro - Motivational Slides on board)

Week 1-2

Gözde ÜNAL

ML Review

Machine Learning Review:

SL example [ex: Digit recognition]

Input $\left\{ \boxed{2}\ 2,\ \boxed{0}\ 0,\ \boxed{5}\ 5, \ \vdots \right\}$ $\longrightarrow$ Machine Learning Algo $\longrightarrow$ $y^{(i)}$ $\vdots$

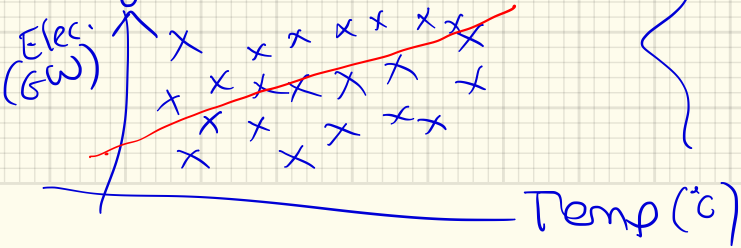
$x^{(i)} \in \mathcal{X}$ $\qquad y^{(i)} \in \mathcal{Y}$

1. hypothesis fn.

$h : \mathcal{X} \to \mathcal{Y}$ $\quad : h(x^{(i)}) \approx y^{(i)}$

Linear Regression: $y \in \mathbb{R}$.

ex Predicting electricity demand:



Dates	Tempe.	power Elec (GW)
01-06.2017	20°c	80
02.06.2017	25°	,
⋮	⋮	⋮

ML Notation: Input features $x^{(i)} \in \mathbb{R}^n$; $i = 1, \dots, m$

ex: Electricity Demand $\approx \theta_1 \cdot \text{Temp} + \theta_2$

eg. $x^{(i)} = \begin{bmatrix} \text{Avg Temp.}^{(i)} \\ 1 \end{bmatrix}$

Output : $y^{(i)} \in \mathbb{R}$ (regression task)

$y^{(i)} \in \{0, 1\}$, $\{0, 1, \dots, k\}$ } classification

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

Model parameters: $\theta \in \mathbb{R}^n$: same size as input features

1 Hypothesis fn : $h_\theta(x) = \underline{\theta}^T \underline{x} = \sum_{j=1}^{n} \theta_j x_j$

2 Loss fn : How do we measure how "good" a hypothesis fn. is

$h_\theta(x^{(i)}) \approx y^{(i)}$

$\ell(h_\theta(x), y)$: should be defined.

eg. For regression task: a common loss fn: $\ell(h_\theta(x), y) = (h_\theta(x) - y)^2$

All ML algorithms: We define the general ML problem.

Given a $(x^{(i)}, y^{(i)})\Big|_{i=1}^{m}$ a set of input features & outputs
#=m

Goal: Find the parameters that minimize the sum of losses:

$$\min_{\theta} \sum_{i=1}^{m} \ell\left(h_\theta(x^{(i)}), y^{(i)}\right)$$

have this form: we just need to specify:

1. What is the hypothesis fn?
2. What is the loss fn?
3. How do you solve the optimization problem?

Least Squares Problem: → Let's formulate w/ this notation

1. Hypothesis function: $h_\theta(x) = \theta^T x$
2. Squared loss func: $\ell(h_\theta(x), y) = (h_\theta(x) - y)^2$
3. w/ 1 & 2 leads to the ML optimiz. problem: →

$$\min_{\theta} \sum_{i=1}^{m} \ell(h_\theta(x^{(i)}), y^{(i)}) = \min_{\theta} \underbrace{\sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})^2}_{\mathcal{L}(\theta)}.$$

$\mathcal{L}(\theta)$: overall obj. fn.

↳ Calculate the Gradient of our objective fn.

$$\nabla_\theta \left(\sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})^2 \right) = \sum_{i=1}^{m} \nabla_\theta (\theta^T x^{(i)} - y^{(i)})^2$$

$$\nabla_\theta \mathcal{L}(\theta) = \sum_{i=1}^{m} 2 (\theta^T x^{(i)} - y^{(i)}) x^{(i)}$$

Gradient Descent (GD):

$$\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta)$$

For the LS problem

GD: $$\theta \leftarrow \theta - \alpha \sum_{i=1}^{m} x^{(i)} (\theta^T x^{(i)} - y^{(i)})$$

1. iterative optimiz.
2. Introduce vector matrix notation

$$\underline{\underline{X}} = \begin{bmatrix} \underline{x^{(1)T}} \\ \underline{x^{(2)T}} \\ \vdots \\ x^{(m)T} \end{bmatrix} \in \mathbb{R}^{m \times n} \qquad \underline{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} \in \mathbb{R}^m$$

$\theta \in \mathbb{R}^n$

write the LS objective : $\sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})^2 = \| \underline{\underline{X}} \, \underline{\theta} - \underline{y} \|_2^2$

Gradient is: $\nabla_{\underline{\theta}} \| \underline{\underline{X}} \underline{\theta} - \underline{y} \|_2^2 =$

$$\mathcal{L}(\underline{\theta}) = \langle X\underline{\theta} - y , X\theta - y \rangle = (X\theta - y)^T (X\theta - y)$$

$$\nabla_\theta \rightarrow \mathcal{L}(\theta) = \theta^T X^T X \theta - \underbrace{2\theta^T X^T y}_{y^T X \theta} + y^T y$$

$$\nabla_\theta \mathcal{L}(\theta) = 2 \underline{\underline{X}}^T \underline{\underline{X}} \underline{\theta} - 2 \underline{y}^T \underline{X}$$

$$\nabla_\theta \mathcal{L}(\theta) = 0 \Rightarrow X^T X \theta = \cancel{y^T_{m\times 1} X_{m\times n}} \; X^T_{m\times n} y_{m\times 1}$$

$$\Rightarrow \boxed{\theta^* = (X^T X)^{-1} \underbrace{X^T y}}$$

Normal equations

e.g. For the electricity prediction ex; solving this eqn gives θ_1, θ_2,

i.e. red line in the plot on page 1.