CS 725

Recap : Algorithm for Gradient Descent

INPUTS: Initial parameter vector w_0, learning rate $\eta > 0$ (or step size), threshold $\epsilon > 0$, maximum # of iterations N_{max}

OUTPUT: final parameter vector w^*

$w \leftarrow w_0$

for $t = 1 \ldots N_{max}$

Compute $\nabla_w L(w, \mathcal{D})$

$$\nabla_w L(w,\mathcal{D}) = \nabla_w \frac{1}{n} \sum_i \big(y_i - \underbrace{w^T x_i}_{\hat{y}_i}\big)^2$$

$$= \frac{2}{n} \sum_i (\hat{y}_i - y_i) x_i$$

EXPENSIVE ∵ computation involves all n points

$w \leftarrow w - \eta \nabla_w L$

if $(\|\nabla_w L\|_2^2 \leq \epsilon)$, then break;

// Alternate: if $(\|w_{t+1} - w_t\|_2^2 \leq \epsilon)$

return w

VARIANTS OF GRADIENT DESCENT

- Full GD : Wt update rule is $w \leftarrow w - \eta \left(\nabla_w L(w, \mathcal{D}_{train}) \right)$
- Stochastic GD (SGD) : wt update rule is $w \leftarrow w - \eta \left(\nabla_w L(w, \mathcal{D}_{random}) \right)$
  where $\mathcal{D}_{random} = \{(x_i, y_i)\}$ where (x_i, y_i) is picked at random from $\mathcal{D}_{train}$
- (MINI) BATCH GD : Wt update rule is $w \leftarrow w - \eta \left(\nabla_w L(w, \mathcal{D}_{batch}) \right)$
  where $\mathcal{D}_{batch} = \{(x_i, y_i)\}_{i=1}^{B}$, where B is batch size
  $B = |\mathcal{D}_{batch}|$

Definitions

TRAINING SET : $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i, y_i are i.i.d. (independent & identically distributed) points

DEVELOPMENT OR VALIDATION SET (HELDOUT SET) : $\mathcal{D}_{val}$ is used to tune hyperparameters

EVALUATION OR TEST SET : $\mathcal{D}_{test}$; not to be used for hyperparam tuning

WEIGHTS or PARAMETERS : w

HYPERPARAMETERS : Predefined values for a chosen model. E.g. η, B, N_{maxi}, ϵ, σ, initial wt vector w_0

TRAIN/DEV/TEST ERROR : $\frac{1}{n} \sum_i (\hat{y}_i - y_i)^2$

PROBABILISTIC MODEL OF LINEAR REGRESSION

Consider $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$. Let each y_i be a noisy target value defined as: $y_i = f(x_i) + \epsilon_i = \omega^T x_i + \epsilon_i$

Where ϵ_i is the noise in the target values that lead to uncertainty in the model.

Let ϵ_i be drawn from a Gaussian distribution with mean 0 and variance σ^2

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$y_i = w^T x_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \qquad \left[\begin{array}{c} Cov(\epsilon_i, \epsilon_j) = 0 \\ \text{if } i \neq j \end{array} \right]$$

$$\Updownarrow$$

$$y_i \sim \mathcal{N}(w^T x_i, \sigma^2)$$

$$\Updownarrow$$

$$P(y_i \mid x_i, w) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right\}$$

$$\underbrace{P(y_1, y_2, \ldots, y_n \mid x_1, x_2, \ldots, x_n, w) = \prod_{i=1}^{n} P(y_i \mid x_i, w)}_{\text{(CONDITIONAL) LIKELIHOOD OF DATA}}$$

Find w that maximizes the likelihood of observed data $\mathcal{D}$:

This is MAXIMUM LIKELIHOOD ESTIMATION (MLE)

[frequentist view of ML]

$$W_{MLE} = \arg\max_{w} \left(\prod_{i} P(y_i | x_i, w) \right)$$

$$= \underbrace{\arg\max_{w} \sum_{i} \log P(y_i | x_i, w)}_{\text{LOG LIKELIHOOD FUNCTION}}$$

MOTIVATE MLE w/ A COIN TOSSING EXAMPLE

Say we want to estimate the prob of a coin landing on heads, denoted by θ. Consider a dataset of N coin tosses, with N^H heads and N^T tails. What is the max. likelihood estimate of θ?

Soln:

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}\, P(\mathcal{D}|\theta) = \underset{\theta}{\operatorname{argmax}}\, \theta^{N^H}(1-\theta)^{N^T}$$

$$= \underset{\theta}{\operatorname{argmax}}\, \underbrace{N^H \log\theta + N^T \log(1-\theta)}_{LL(\theta)}$$

$$\frac{\partial LL(\theta)}{\partial \theta} = 0 \Rightarrow \frac{N_H}{\theta} - \frac{N_T}{(1-\theta)} = 0$$

$$\Rightarrow \theta_{MLE} = \frac{N_H}{N_H + N_T} = \frac{N_H}{N}$$

Q. Say $\theta \in \{0.4, 0.6\}$

Data: 3 coin tosses, 2H, 1T

What is θ_{MLE}? $\underline{0.6}$

What is MLE for linear regression?

$$w_{MLE} = \underset{w}{\operatorname{argmax}} \sum_i \log P(y_i \mid x_i, w)$$

$$= \underset{w}{\operatorname{argmax}} \sum_i \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right\} \right)$$

$$= \underset{w}{\operatorname{argmax}} \; K - \frac{1}{2\sigma^2} \sum_i (y_i - w^T x_i)^2$$

$$= \underset{w}{\operatorname{argmin}} (y_i - w^T x_i)^2$$

LEAST SQUARES ERROR