Week: 1+2 | Reg+Opt

week1

Regression

Either find a relationship between X and Y (Interpretation)
or try to train a model to predict Y (prediction)

Linear Regression

$$f_w(x) = x^T w \quad \text{(Lin. Reg.)}$$

cost function

$$L(w) = \begin{cases} MSE: \ L_{(w)} = \frac{1}{2N}\sum_{n=1}^{N}(y_n - x_n^T w)^2 \\ \text{(Lin. Reg.)} \\ MAE: \ L_{(w)} = \frac{1}{N}\sum_{1}^{N}|y_n - x_n^T w| \end{cases} \quad \left(= \frac{1}{2N}(y - Xw)^T(y - Xw)\right)$$

week2

Optimization:

find best w: $w^* \Rightarrow$ GridSearch (not guaranteed, granularity problem)
$\Rightarrow$ Gradients

Lin. $L(w) = \frac{1}{2N}\|y - X^T w\|^2 = \frac{1}{2N} e^T e \rightarrow \nabla L(w) = -\frac{1}{N} X^T e$

Gradient descent (GD)

$$w^{t+1} = w^t - \gamma \nabla L(w^{(t)}) \quad (\gamma: \text{Learning rate})$$

Stochastic Gradient descent (SGD)

$$w^{(t+1)} = w^t - \gamma \nabla L_n(w^{(t)}) \quad \text{(sample one point } n \text{ at random)}$$

Mini-Batch Stochastic Grad. desc.

(GPU parallelizable...)

$$g = \frac{1}{|B|}\sum_{n \in B} \nabla L_n(w) \rightarrow w^{(t+1)} = w^t - \gamma g$$

Subgradients:

(non-smooth opt.)
(if L is not diff. and not necessarily convex)

if $L(w) = h(q(w))$ | i.e. $L(w) = \frac{1}{N}\sum |y_n - x_n^T w|$
where h is not differentiable and q is
Then $g \in \partial h(q(w)) \cdot \nabla q(w)$ is a subgradient for $L(w)$ at w

and $\rightarrow$ GD & SGD with g: $w^{(t+1)} = w^t - \gamma g$

Costs (Lin. Reg.: $y = x^T w$)

GD: $O(N \cdot D)$

SGD: $O(D)$

1 step of (S)GD (Mini-batch)
$O(|B| \cdot D)$ $\begin{cases} B = N \ (GD) \\ B = 1 \ (SGD) \\ B \ (Mini\text{-}B\ SGD) \end{cases}$

$L(w) = \sum_n L_n(w)$

γ: Learning rate
$\gamma = c/t$ (Const)
Stopping criterion: $\|\nabla L(w)\| \ll K$ small K
Convergence: is only guaranteed when $\forall \gamma < \gamma_{min}$ (γ_{min} found depend on the problem)