Lec 13 Linear / Ridge Regression

Recap.

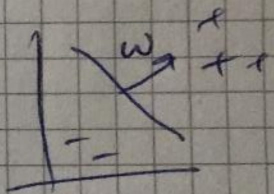
we have $P(y|x) = \frac{1}{1+e^{-w^T x y}}$

logistic loss function - loose approx. of "zero-one loss".

$$\boxed{l(w) = \sum_{i=1}^{n} \log\left(1+e^{-w^T x_i y_i}\right)}$$

zero-one loss, - go over the dataset and just count how many I get wrong. Not differentiable nor continuous so we can't optimize it.

$\mathcal{L}(x_i, y_i, w) = \log\left(1+e^{-w^T x_i y_i}\right)$



Classific.-n: correct $\Rightarrow$ w is +ve $\Rightarrow$ $w^T x_i$ is +ve

$\Rightarrow e^{-(\text{positive qtty})} \Rightarrow 0$

$\mathcal{L}(x_i, y_i, w) \approx \log(1) = 0$

wrong $\mathcal{L}(x_i, y_i, w) =$

$\hookrightarrow$ y is +ve, ~~but~~ assuming we classify a positive point as -ve $\Rightarrow w^T x_i$ is -ve $\Rightarrow e^{-(\text{negative qtty})}$

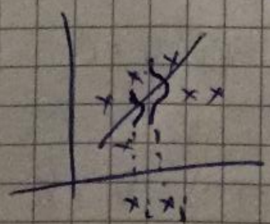
$\Rightarrow e^{+(>0)} \Rightarrow$. Now $1+e^{\text{large}} \approx e^{\text{large}}$

so $\mathcal{L}(x_i, y_i, w) = -w^T x_i y_i \Rightarrow$ large loss

---

OLS (Ordinary Least Squares)

$y_i \in \mathbb{R}$, Assumption for each x, there is a gaussian distribution of possible y values



i.e. $y_i = w^T x_i + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$

or $y_i \sim N(w^T x_i, \sigma^2)$

Now we have.

$$P(y_i \mid \vec{x}_i, w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(w^T x_i - y_i)^2}{2\sigma^2}}$$

Maximize for w $\Big\langle$ MLE / MAP

MLE: $\operatorname{argmax}_w \prod_{i=1}^{n} P(y_i \mid \vec{x}_i, w)$

take log $\quad = \operatorname{argmax}_w \left[\underbrace{\sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)}_{\text{constant}} + \underbrace{\left(-\frac{1}{2\sigma^2}\right)}_{\text{constant}} \sum_{i=1}^{n} (w^T x_i - y_i)^2 \right]$

$$= \operatorname{argmin}_w \sum_{i=1}^{n} (w^T x_i - y_i)^2$$ (multiply by -1) to make min.

average $\quad = \operatorname{argmin}_w \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2$

MAP:

$$P(w \mid \underbrace{y_i, x_i}_{D}) = \frac{P(D \mid w)\, P(w)}{Z}$$

$P(w)$ prior

$w \sim N(0, \tau^2)$

$P(w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{w^2}{2\tau}}$

we do,

$$\operatorname{argmax}_w P(y_1, x_1, \ldots, y_n, x_n \mid w)\, P(w)$$ drop Z

??

x_i is on other side of "|" as w is independent of x_i

$$= \operatorname{argmax}_w \prod_{i=1}^{n} P(y_i \mid x_i, w)$$

$$= \operatorname{argmax}_w \sum_{i=1}^{n} \log\left(\underbrace{P(y_i \mid x_i, w)}_{\approx \text{MLE}} P(w)\right)$$

$$= \operatorname{argmin}_w \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[(w^T x_i - y_i)^2 + \frac{1}{2\tau^2} w^T w \right]$$ $w^T w = w^2$

$$= \operatorname{argmin}_w \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[(w^T x_i - y_i)^2 \right] + \frac{n}{2\tau^2} w^T w$$

$$= \operatorname{argmin}_w \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_2^2$$