Linear Models

Tuesday, January 28, 2020 13:10

Grad-School Application Analysis
- Approve/Deny — Perceptron — Classification (+1/-1) error, PLA, pocket alg
- Amount of Scholarships — Linear Regression — Squared Error, Pseudo-inverse
- Probability of acceptance — Logistic Regression — cross entropy, gradient-descent

Hybrid: logistic reg + threshold ½ ⇒ classification

$$\text{sign}(\underline{w}_1^T\underline{x}) = \text{sign}(\theta(\underline{w}_2^T\underline{x}) - \tfrac{1}{2})$$

often done

Differences ?

⇒ the weights would be different: $(\underline{w}_1^* \neq \underline{w}_2^*)$

PLA vs. Grad Des.

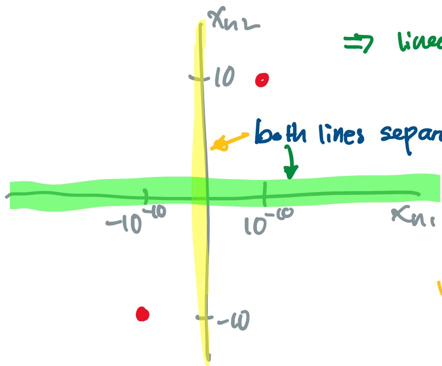
+1/-1 error vs. cross entropy error.

ex. $d=2$ $N=2$ $\underline{x}_1 = (1, 10^{-10}, 10)^T$, $y_1 = +1$

$\underline{x}_2 = (1, -10^{-10}, -10)^T$ $y_2 = -1$

$d=2$

Linear Classification:



⇒ linearly separable

both lines separates the data: both has $\underline{w}^T\underline{x}_n \cdot y_n > 0$

$$\Rightarrow E_{in}(\underline{w}) = \frac{1}{2}\sum_{n=1}^{2} \mathbb{1}\{y_n \neq \text{sign}(\underline{w}^T\underline{x})\} = 0$$

yellow line: $\underline{w}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ green line: $\underline{w}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

Logistic Regression:

$$E_{in}(\underline{w}) = \frac{1}{2}\sum_{n=1}^{2} \log(1 + e^{-y_n\underline{w}^T\underline{x}})$$

yellow line: $= \log(1 + e^{-10^{-10}}) \approx$

green line: $= \log(1 + e^{10^{-10}}) \approx$

green line: $= \log(1+e \quad) \simeq$

Using this, we see that second solution is better

Multiclass Logistic Regression

alternative to multiclass binary classifiers:

→ softmax or multinomial logistic regression

$y \in \{1, 2, \dots K\}$

$$w = \begin{bmatrix} \underline{w}_1^T \\ \underline{w}_2^T \\ \vdots \\ \underline{w}_K^T \end{bmatrix} \qquad \hat{p}_w(c \mid \underline{x}) \simeq e^{\underline{w}_c^T \underline{x}}$$

$$\Rightarrow \hat{p}_w(c \mid \underline{x}) = \frac{e^{\underline{w}_c^T \underline{x}}}{\sum_{i=1}^{K} e^{\underline{w}_i^T \underline{x}}} \Bigg\} \text{ softmax classifier.}$$

Loss resembles CE loss:

$$E_{in} = \frac{1}{N} \sum_{n=1}^{N} -\log\left(\hat{p}_w(y_n \mid \underline{x}_n)\right)$$

$$= \frac{1}{N} \sum_{n=1}^{N} \left[\underbrace{\sum_{i=1}^{K} 1\{y_n = i\}}_{y_{ni} \text{ (one-hot)}} \cdot \frac{e^{\underline{w}_i^T \underline{x}_n}}{\sum_{k=1}^{K} e^{\underline{w}_k^T \underline{x}_n}} \right]$$