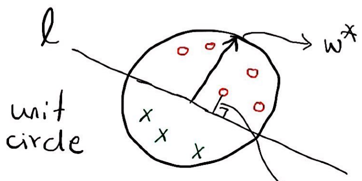
Kimia Noorbakhsh

1. if the data is linearly separable, then there exists a $w^*$ such that $\forall (x_i, y_i) \in D : \underline{y_i (x^T w^*) > 0}$

For convinience, supposethat we re_scale our data such that $\underline{\|w^*\| = 1}$, and $\underline{\|x_i\| \leqslant 1 \ \forall x_i \in D}$




$K = \min_{D} |x_i^T w^*|$

we define K as above. $\left(\begin{array}{c}\text{min distance of } \ell \text{ to} \\ \text{the data points}\end{array}\right)$

$\Rightarrow$ claim : perceptrol algorithm, make at most $1/K^2$ mistakes untit convergence.

Proof of the claim:

from the algorithm definition we have:

$$\begin{cases} y(x^T w) < 0 \rightsquigarrow \text{we need update} \\ y(x^T w^*) > 0 \rightsquigarrow w^* \text{ makes a separating line.} \end{cases}$$

consider on update : $w \longrightarrow w + yx$ ⊛

$$⊛ \Rightarrow (w+yx)^T w^* = w^T w^* + \underline{y(x^T w^*)} \geq w^T w^* + \underline{K}$$

$$\left[ \text{because } \quad y(x^T w^*) = |x^T w^*| \geq K \right]$$

$\Rightarrow$ after each update $w^T w^*$ increases at least $K$.

①

$$⊛ \Rightarrow (w+yx)^T (w+yx) = w^T w + \underbrace{2y(w^T x)}_{<0} + y^2 (x^T x)$$

[because we have update]

$$\xrightarrow{y^2 = 1 \text{ and } \|x\| \leq 1} (w+yx)^T (w+yx) \leq w^T w + 1$$

$\Rightarrow$ after each update $w^T w$ increases at most 1

②

from ① and ②, after $i$ updates we have :

$$\begin{cases} w^T w^* \geq iK \\ w^T w \leq i \end{cases}$$

$\Rightarrow i K \leq w^T w^* = \|w\| \cos\theta$ [$\theta$ is angle between $w$ and $w^*$]

$\leq \|w\| = \sqrt{w^T w} \leq \sqrt{i}$

$\Rightarrow i.K \leq \sqrt{i} \Rightarrow \boxed{i \leq \frac{1}{K^2}}$

$\Rightarrow$ number of updates is at most $\frac{1}{K^2}$

$\Rightarrow$ we will converge finally ✓

---

2. suppose we have $\|\theta_{MAP}\|_2 > \|\theta_{ML}\|_2$

and we know that :

$$\mathbb{P}(\theta_{MAP}) = \frac{1}{(2\pi)^{\frac{n+1}{2}} \times \sqrt{|\tau^2 I|}} \times \exp\left(\frac{-1}{2\tau^2} (\|\theta_{MAP}\|_2)^2\right)$$

by assuption $< \frac{1}{(2\pi)^{\frac{n+1}{2}} \sqrt{|\tau^2 I|}} \exp\left(\frac{-1}{2\tau^2} (\|\theta_{ML}\|_2)^2\right)$

$= \mathbb{P}(\theta_{ML})$

$$\Rightarrow P(\theta_{MAP}) \prod_{i=1}^{m} P(y^{(i)} | x^{(i)}, \theta_{MAP}) \leq P(\theta_{ML}) \prod_{i=1}^{m} P(y^{(i)} | x^{(i)}; \theta_{MAP})$$

$$\begin{bmatrix} \text{because } \theta_{ML} \text{ maximizes} \\ \prod_{i=1}^{m} P(y^i | x^i, \theta) \end{bmatrix} \leq \underbrace{P(\theta_{ML}) \prod_{i=1}^{m} P(y^{(i)} | x^{(i)}, \theta_{ML})}_{\circledast}$$

but $\circledast$ is contradiction, because $\theta_{MAP}$ maximizes $P(\theta) \prod_{i=1}^{m} P(y^i | x^i, \theta)$ but in $\circledast$ it is less than same equation for $\theta_{ML} \rightarrow \times$.

therefor $\|\theta_{MAP}\|_2 \leq \|\theta_{ML}\|_2$ ✓

---

3. claim: binary naive bayes is a linear classifier.

proof: assume that each sample is a d-dimentional vector like: $X = (x_1, x_2, \dots, x_d)^T$.

Assume that the features $x_j$ are also binay: $x_j \in \{0, 1\}$

then if $\mathbb{P}(y=1 | X) \geq P(y=0 | X) \Rightarrow$ predict 1

if $\mathbb{P}(y=0 | X) > \mathbb{P}(y=1 | X) \Rightarrow$ predict 0

By Bayes rule, if $\frac{P(X|y=1)P(y=1)}{P(X|y=0)P(y=0)} \geq 1$ we will predict 1 for the label.

$\circledast$

Because we have the Naive Bayes assumption, we have $P(X|y) = \prod_{i=0}^{d} P(x_i|y)$. So if we re-write $\circledast$, we have:

$$\frac{P(y=1)}{P(y=0)} \times \prod_{i=0}^{d} \frac{P(x_i|y=1)}{P(x_i|y=0)} \geq 1$$

$\circledast^2$

Let $P(y=1) = p$ and $P(x_i=1|y=1) = q_i$

$P(y=0) = 1-p$ $\quad P(x_i=1|y=0) = q'_i$

$$\Rightarrow P(x_i|y=1) = q_i^{x_i}(1-q_i)^{1-x_i}$$

$$P(x_i|y=0) = q_i'^{x_i}(1-q'_i)^{1-x_i}$$

$$\Rightarrow \circledast^2 = \frac{p}{1-p} \times \prod_{i=0}^{d} \frac{q_i^{x_i}(1-q_i)^{1-x_i}}{q_i'^{x_i}(1-q'_i)^{1-x_i}} \geq 1$$

$\Longleftrightarrow$

$$\left(\frac{p}{1-p} \times \prod_{i=0}^{d} \frac{1-q_i}{1-q'_i}\right) \cdot \prod_{i=0}^{d} \left(\frac{q_i}{q'_i} \cdot \frac{1-q'_i}{1-q_i}\right)^{x_i} \geq 1$$

$\overset{\log}{\Longleftrightarrow}$

$$\text{const} + \sum_{i=0}^{d} x_i \log\left(\frac{q_i}{q'_i} \cdot \frac{1-q'_i}{1-q_i}\right) \geq 0$$

now let $w_i = \log\left(\frac{q_i}{q_i'} \cdot \frac{1-q_i'}{1-q_i}\right)$

$\Rightarrow$ we have : $\text{const} + \sum_{i=0}^{d} x_i w_i \geq 0$ **

Therefore if ** holds, we will predict label 1 $\Longrightarrow$ The classifier is linear ✓

---

4. define : $\varepsilon_i(x) = y_i(x) - h(x)$

Then we would have :

$$E_x[(H_M(x) - h(x))^2] = E_x\left[\left(\frac{1}{M}\sum_{i=1}^{M} y_i(x) - h(x)\right)^2\right]$$

$$= E_x\left[\left(\frac{1}{M}\sum_{i=1}^{M} \varepsilon_i(x)\right)^2\right]$$

$$= \frac{1}{M^2} E_x\left[\left(\sum_{i=1}^{M} \varepsilon_i(x)\right)^2\right] \quad *$$

by cauchy inequality we know that :

$$(1^2 + 1^2 + \cdots + 1^2)(\varepsilon_1^2(x) + \cdots + \varepsilon_M(x)^2) \geq \left(\sum_{i=1}^{M} \varepsilon_i\right)^2$$

Therefore: $m \sum_{i=1}^{m} \varepsilon_i^2(x) \geq \left( \sum_{i=1}^{m} \varepsilon_i(x) \right)^2$

combining with ⊛, we have:

$$\frac{1}{M^2} \underset{x}{E}\left[ \left( \sum_{i=1}^{M} \varepsilon_i(x) \right)^2 \right] \leq \frac{M}{M^2} \underset{x}{E}\left[ \sum_{i=1}^{M} \varepsilon_i^2(x) \right]$$

$$= \frac{1}{M} \sum_{i=1}^{M} E_x\left[ (y_i(x) - h(x))^2 \right]$$

✓

---

5. a) if we have M features, then we would at most $2^M$ different data points, and a decision tree of depth $m$, has at most $2^m$ leaves which covers all possible data. Therefore we can classify our dataset with <u>zero</u> training error.

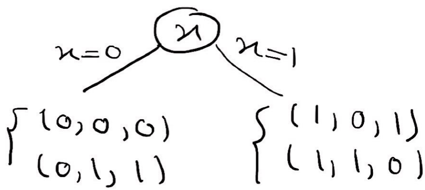
b) No, consider this counterexample:
suppose we have $2 = m$ featurs, $x$ and $y$ which each can be $\underline{0}$ or $\underline{1}$.

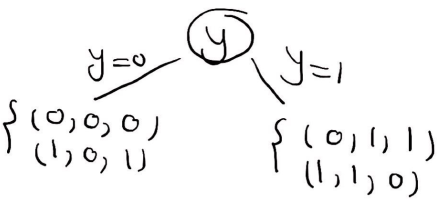
now suppose we have the following data:

| x | y | label |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |

$\longrightarrow$ (0, 0, 0), (0, 1, 1), (1, 0, 1), (1, 1, 0)

For a tree with depth $m-1=1$, we have these two cases:




we can see that in both choices for root node we have 50% accuracy.