Naive Bayes-02
http://youtu.be/...

Q: Guess the word?

"Bag of words"

Agenda:

1. Laplace Smoothing
2. Hyper parameter
3. Bias Variance Tradeoff
4. Imbalanced data
5. Underflow Problem
6. Feature Importance
7. Impact of outliers

Laplace Smoothing.

Review

Spam → buy, viagra, ..., free

$P(w_1|spam)$
$P(w_2|spam)$

$w_1, w_2, w_3, \ldots, w_d$

$w' \rightarrow$ new word

$w' \notin \{w_1, w_2, w_3, \ldots, w_d\}$

$P(w_1|spam), P(w_2|spam), \ldots, P(w_d|spam)$

$\Rightarrow P(w_1|spam), P(w_2|spam), \ldots, P(w_d|spam), \boxed{P(w'|spam)}$

$P(w'|spam) = 0$

$P(w_1|ham), P(w_2|ham), \ldots, P(w_d|ham), \boxed{P(w'|ham)}$

$P(w'|ham) = 0$

Problem ?

w' can figure in ?

$P(w'|y=1) = \frac{1}{\ldots}$ spam mail from us

$P(w'|y=0) = 0$

Laplace Smoothing

$$P(w'|y=1) = \frac{n_{w'}}{n_1}$$

$n = 0$

$$\Rightarrow \frac{n_{w'}+\alpha}{n_1+\alpha C}$$

α → pseudo value → 1

C → (unique) words → 2

Hyper Param.

Bias Variance tradeoff.

Review $\boxed{K=1}$

$$P(w'|y=1) = \frac{n_{w'}+1}{n_1+2}$$

$n_{w'}=100$
$n_1=1000$

$$\Rightarrow \frac{1}{\ldots} \approx 0.098$$

$$P(w'|y=0) = \frac{n_{w'}+1}{n_0+2} \Rightarrow \frac{0+1}{\ldots} \approx \ldots$$

$n_{w'}=0$
$n_0=1000$

≈ 0.0008

Case of $K=0$: not modeling baseline.

$n_{w'}=100$, from $y=0$

$$P(\text{ham}|w'=1) = \frac{n_{w'}}{n_1+n_0} = \frac{100}{1000} \approx 0.1$$

Case of $K=1$

$n_{w'}=100$, from $y=0$

$$P(\text{ham}|w'=1) = \frac{100+1}{1000+2} \approx \frac{101}{\ldots} \approx 0.1$$

Case of $K=\infty$

$$P(\text{ham}|w'=1) = \frac{n_{w'}+\alpha}{n_0+\alpha C} = \frac{100+\alpha}{1000+\alpha C} \approx 0.08 \approx 0.0$$

Overfitting

$$P(\text{ham}|w'=0) = \frac{0+\alpha}{1000+\alpha C} \approx 0.0$$

All are good Predictions ≈ 0.0

① $P(w_1|y_k) = \prod_{i=1}^{d} P(w_i|y=k)$, $P(y=1)$

② $P(y=k|x) = \prod_{i=1}^{d} P(w_i|y=k)$, $P(y=0)$

$P(y=0|x) = 0.02$

$P(y=1|x) = 0.03$

$\alpha \uparrow \longrightarrow$ underfit

$\alpha \downarrow \longrightarrow$ overfit

How to pick α ?

Cross Validation

Hyper Param tuning

Grid Search

Random search

Imbalanced Data.

$P(y=1) = 0.3$
$P(y=0) = 0.7$

→ underfit → Predict bundle Ham.

Up sampling

Down sampling

Use weighing, not to remove data

SMOTE (Synthetic Minority Over Sampling Technique)

Xgboost, "from long ago"

$\prod_{i=1}^{n} P(w_i|y=1) \approx 0.2$

$\prod_{i=1}^{n} P(w_i|y=0) \approx 0.2$

$P(y=1|x) \approx 0.2 \times 0.3 \Rightarrow 0.06$

$P(y=0|x) \approx 0.2 \times 0.7 \Rightarrow 0.06$

Function made first → equal = 0.5

$0.2 \times 0.3 \approx 0.06$

$0.2 \times 0.7 \approx 0.05$

$\frac{0.06 \times 0.3}{\frac{0.06 \times 0.3}{0.06 \times 0.3 + 0.05 \times 0.7}} \approx 0.4$

Underflow Problem.

$$P(w_1|y_k) \Rightarrow \prod_{i=1}^{d} \underbrace{P(w_i|y_k)}_{0<1} \cdot \underbrace{P(y=k)}_{0<1}$$

$d \uparrow$

$P(w_i) \rightarrow [0, 1]$ Below

$\Rightarrow$ $0.1 \times 0.2 \times 0.3 \times 0.4 \times \ldots$

$\Rightarrow$ Very very small number

$\Rightarrow$ $\underbrace{\text{underflow}}_{\text{Computer}}$ ≈ 0

Return, from Function

$$P(w_1|y_k) = P(w_1|y_k) \cdot \prod_{i=2}^{d} P(w_i|y_k)$$

Mainly for our benefit

$$\log\left[P(w_1|y_k)\right] \Rightarrow \log\left[P(w_1|y_k)\right] + \left[\log P(w_2|y_k) + \log P(w_3|y_k) \ldots\right]$$

$\log[0.0001]$

$$\Rightarrow \underbrace{\log\left[P(w_1|y_k)\right] + \sum_{i=2}^{d} \log P(w_i|y_k)}$$

if $\log(y_k) > \log(y_l)$

then $y_k > y_l$

Feature Importance

$w_1 \rightarrow P(w_1|y=1)$
$\rightarrow P(w_1|y=0)$

"Large" → $P(w_1|y=1) \uparrow$
$\rightarrow P(w_1|y=0) \downarrow$

"Small" → $P(w_1|y=1) \downarrow$
$\rightarrow P(w_1|y=0) \uparrow$

$y_q = w_1 \text{ or } w_2 \text{ or } w_3 \text{ or } w_4$

$y_q = 2$

Importance

$P(w_1|y=1) \approx 0.8$

Impact of outliers

Review

$y_q = w_1 \text{ or } w_2 \text{ or } w_3 \text{ or } w_4$ → 4 Classes

Outliers

Outliers → "Rare words"

Outliers → "Rare words"

$P(w_1|y=0) = \frac{0+\alpha}{\text{Total}+\alpha C}$

αC for α → Laplace Smoothing

α → pseudo value → Instead of adding just 1 number

Place α → Similar value of α

$\{\alpha$ for α ; α ; α ;