X — Y

X ↓ cate ↓

Y ↓ cate

X = 2 cate

Y = 2 cate

→ Topic 2

3 cate → { plot
contingency table
not OR (OR is only for 2 × 2 table).

Pr (large | M) = ≈ 3.7%

Pr (large | f) = ~ 2.8%.

} conditional prop / prob could help.

---

order size

regressor: income

	large	small	conditional pro
low	→	→	100%
med			100%
high			100%

Prop (large | low) = 5%

Prop (large | med) = 10%

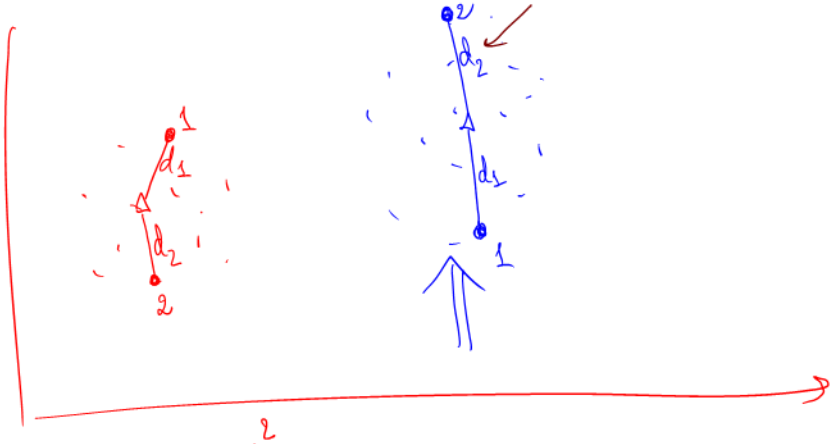
Prop (large | high) = 20%

X * Y

3 x 2

a x b

→ KNN ←
→ DT ←
→ NB ←
→ LO.R ← ←
LM



SS for blues:

d_1^2 : point 1 → Δ

d_2^2 : point 2 → Δ

⋮

all blue points: $\sum_{i \in \text{blue}} d_i^2$

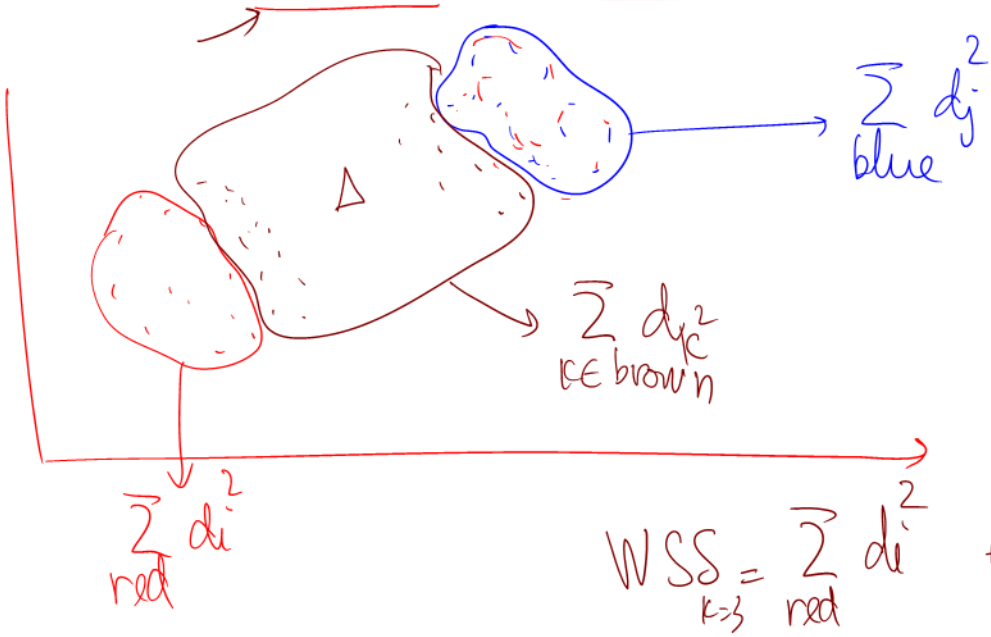
SS for reds: d_1^2

d_2^2

⋮

all red points ⇒ $\sum_{j \in \text{reds}} d_j^2$

$$WSS_{k=2} = \sum_{\text{blue}} d_i^2 + \sum_{\text{red}} d_j^2$$



$\sum_{\text{red}} d_i^2$

$$WSS_{k=3} = \sum_{\text{red}} d_i^2 + \sum_{\text{blue}} d_j^2 + \sum_{\text{brown}} d_k^2$$

Compare: $WSS_{k=2}$ vs $WSS_{k=3}$

value of k that gives smaller WSS is better. However we would prefer smaller k.

hdb : try with $k = 1 \rightarrow WSS_1$

$k = 2 \Rightarrow WSS_2$

$k = 10 \rightarrow WSS_{10}$.

$\sim x_1 + x_2$ linear

$\log \frac{p}{1-p} = \cdots\cdots$ logistic $\rightarrow$ Y = response $\begin{cases} 0 \\ 1 \end{cases}$

$p = \text{Prob}(Y = 1)$

---

to find best k for KNN:

fit KNN with $k = 1$ for train data (80%) $\rightarrow$ check gof. $\rightarrow$ test set

fit 2-NN for train data (80%) $\rightarrow$ check g.o.f. $\rightarrow$ for test set

N fold for 80% (traindata).

N folds $\Rightarrow$ find best k

$\rightarrow$ check for test set.