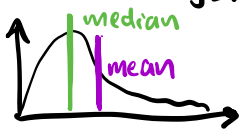
UGBA 104 FINAL

Frequency Distributions

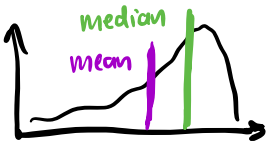
Variables

	Customer #	Purchase Item	Quantity
unique observations	1	A	10
	2	B	15
	⋮		

Histogram Skewness (hint: L/R skew is the opposite of what they look like intuitively).



right skew



left skew

Binning Guide for Quantitative Data
- generally 15-20 bins
- needs to be ME & CE
- Apprx bin width = $\frac{\text{Largest Val} - \text{Smallest Val}}{\text{\# des. bins}}$

CROSS-TABULATION

Normal Distribution: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-M}{\sigma}\right)^2}$

Standardization of Data: $x_{std} = \frac{x - M}{\sigma}$

AKA: "z-score": the # of stan. devs away from the mean a datapoint is.

σ: standard dev.
M: mean / expected value

"outlier" ≙ Any datapoint with a z-score outside of [-3, 3]

Essence of cross-Tab: slice & dice data into segments & analyze individually

Simpson's paradox: Trends in the "sliced & diced" data may disappear in whole dataset

CLUSTERING

Euclidean Distance: how we measure "similarity" between observations:

$duv = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2}$

$duv_{stan} = \sqrt{\sum_{i=1}^{n} (z_{ui} - z_{vi})^2}$ → standardized duv is better otherwise scale of raw data is misleading.

K-means Clustering INPUT: # of clusters K → OUTPUT: cluster assignments.

1. Assign clusters randomly
2. Find centroids of each cluster
3. Reassign to closest centroid
4. Re-calculate centroid location
5. Repeat until state convergence

Notes:
- converges by definition
- does not always reach global optimum
- "UNSUPERVISED" ML

Business context consid.
- what variables are relevant for the business question?
- what clusters make sense?
- what clusters align w/ expert domain knowledge?

ASSOCIATION RULES

Rule ≙ If {Antecedent (A)} then {consequent (C)}

Support Count ≙ # of transactions where {A, C} occurs = count(A & C)

Support ≙ % of transactions that include item set {A, C} "X % of data exhibits A, & C"

$= P(A \& C) =$ support count / tot. count.

Confidence ≙ $P(C|A) = \frac{P(C \& A)}{P(A)}$ = support of rule / support of A "X% of data that has A also has C"

// can be misleading if consequent C is too frequent.

Lift Ratio ≙ $\frac{P(C|A)}{P(C)} = \frac{P(C \& A)}{P(C) * P(A)} = \frac{\text{confidence of rule}}{\text{support of consequent}}$

Lift Ratio	Interpretation
= 1	A & C are statistically independent
> 1	"Positive Association between A&C" → Rule is better than guess
< 1	"Negative Assc. btw. A & C" → Rule is worse than guess

Properties

① Comparing 2 rules, it's poss. to have ↑ confidence but a ↓ lift ratio
② if {A} then {B} & if {B} then {A} will have the SAME lift ratio

What Rules should we go after

① Common (↑ support)
② Reliable (↑ confidence)
③ Better than guess (Lift Ratio > 1)
④ Actionable
⑤ Non-obvious

BINARY CLASSIFICATION

Input Variables → Output Variable

Classification vs Regression.

Confusion Matrix

Actual Class \ Predicted Class	0	1
0	TN (n_{00})	FP (n_{01})
1	FN (n_{10})	TP (n_{11})

Performance Measures

- Accuracy = $\frac{n_{00} + n_{11}}{n_{00} + n_{10} + n_{01} + n_{11}}$
- Overall Error Rate: $\frac{n_{01} + n_{10}}{n_{00} + n_{10} + n_{01} + n_{11}}$

sum to 1

- Sensitivity ⇔ Recall = $\frac{n_{11}}{n_{10} + n_{11}}$
- Specificity = $\frac{n_{00}}{n_{01} + n_{00}}$
- Precision = $n_{11} / (n_{11} + n_{01})$

1 is best

Notes

- false positive ≤ false negative, need to look at context
- Class Imbalance: too FEW "positive" examples to learn from
- DO NOT EVALUATE performance on TRAINING SET.
- Classification is SUPERVISED ML

LOG-REG

Basis Log Reg function = $(1 + e^{-(b_0 + b_1 x_1 + \ldots + b_n x_n)})^{-1} = \hat{p}$

Output: probability $\hat{p}$ $(0 < \hat{p} < 1)$ ≷ Inputs: $(x_1 \ldots x_n)$ ≷ Tune coefficients: $(b_0, b_1, \ldots b_n)$

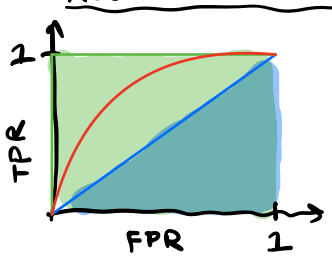
Cutoff probabilities if $\hat{p} > 0.5$ → classify as 1 (0.5 default)

- As cutoff ↓
  - # FP & # TP both increase
  - TP rate & FP rate both incr.
  - TPR ⇔ Sensitivity
  - FPR ⇔ (1 − specificity)

PROS: Interpretable, Give association insights to good predictor vars

CONS: relies on a very specific form, fails to capture complex patterns

ROC: Receiver Operating Characteristics



- Perfect classifier
- Random classifier
- ROC curve

The larger ∫ ROC, the better the classifier

∫ROC = 0.5 ∫ROC = 1

k-NN

if k = 1: too complex, high variance, overfitting
if k = N = # of samples: too general, high bias, underfitting

PRACTICE

Which pair most similar?

(79, 95) A (129, 35) B (88, 148) C

mean var 1: $\frac{79 + 129 + 88}{3} = 98.667$

std var 1: = 26.65 // MATLAB

mean var 2: $92.\overline{6}$

std var 2: 56.53

$d_{AB} = \sqrt{(1.1381 + 0.7379)^2 + (1.02 + 0.0413)^2} = 2.1554$

$d_{AC} = 0.9964$ → shortest dist. A & C are most similar (Ellen & Taylor Swift)

$d_{BC} = 2.522$

Standardized data

A: (−0.7379, 0.0413)
B: (1.1381, −1.02)
C: (−0.4002, 0.9787)

*NOTE!: you don't have to go through standardization if both variables are absolutely most similar! (saves a lot of work)