Classification:

age | tumor size | malignant

predictors / features / characteristics (attributes) | target

↳ relationship between age and tumor size that makes target yes or no (malignant, not malignant)

→ learn model → model:

tumor size > age/2 determines malignant

→ sometimes, based on the data, there are multiple solutions to the "threshold" / prediction
so we need to define how we are going to pick the "threshold" depending on the best we

→ sometimes, there are no correct answer
- → could be because we have wrong or insufficient attributes for the task → wrong: rhinos and elephants using age instead of weight to classify animal species
- → could be because it is a noisy / probabilistic relationship → the problem just doesn't have an exact solution.
  - → the relationship is naturally random
  - → there may be overlap (rhino versus elephant weights)
  - ↳ we can define it so that we only make a reasonable amount of mistakes)

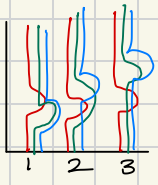
the feasability

Takeways:
1.) many correct answers
2.) no correct
   1.) no relationship
   2.) still useful model if more or less correct

① how do we define these predictors? as good or bad?
② how do we know we've done a good job at classification?

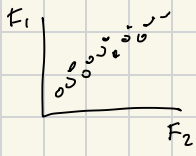
↳ What constitutes a good feature / predictor?

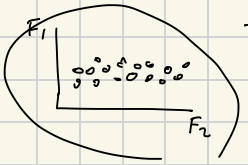
review length



we want to be in the blue situation if we want to be able to classify reviews based on review length

we have feature 1 and 2?
→ do we want them to have a relationship or not?





→ we do not want them to be related at all → as independent as possible

how do we go from one to other?
SVD to select features / reduce but loose interpretability

→ linear or any relationship? we want to avoid all types of relationships.

We want features to be related to the target but not to each other.

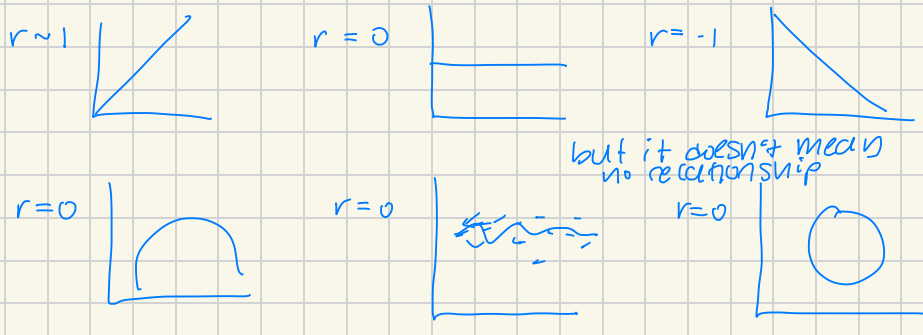
pearson correlation

Correlation:

correlation between X, Y (CRV)

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

- age, tumor size
- → how far away from mean x/y
- → do they move same/different direction
- → covariance – measure of how two variables change together.
- → normalizing for their respective variances → standard dev. of x, y



0 covariance → variables don't have a linear relationship

but it doesn't mean no relationship

some sort of relationship not captured by this correlation. (linear)

What if X is continous, but Y is a class (not a continous variable, but instead {yes or no}) ?

→ Y is {yes, no}, colors, cities, etc. → nominals → no order – need to look at the means of → ANOVA and how they differ across each nominal value of Y

→ Y is {Terrible, Bad, Ok, Good, Great} → Ordinal → natural order to it

→ we can assign numbers to each category like {0, 1, 2, 3, 4} But the difference between Terrible and bad is not the same as ok and bad

Pearson looks to actual values (how far apart BAD and OK are)

Spearman replaces values with ranks (only order matters)

Spearman Coefficient Example

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

1 – and 6 times to make the value mapped between -1 and 1

sum of squared differences is maximized when $\frac{n(n^2-1)}{3}$?

increase X → seems to also increase ordinal variable Y

n = sample size

$d_i = u_i - v_i$ = the difference in the ranks of the two observations

instead of using the exact numbers assigned, we can compare their rank/position in the data so it doesn't matter how far Bad is from Ok, just that Ok comes after Bad

???

Correlation vs. Causation:

Temperature and ice cream sales are positively correlated

1.) but in the desert when there is no ice cream, there is no spike in sales →

2.) ice cream sale increases do not cause the temp to rise.

temperature is not causing ice cream sales just happen simultaneously

→ confounding variable

→ sleeping with shoes on strongly correlated with waking up with a headache.

Simpson's paradox – trend appears in several groups of data, but disappears or reverses when the groups' are combined

Causation → very hard to show things are causally linked through observational data especially if relationship isn't deterministic

→ testing (experiments (especially) with a control group → ethical considerations

having an aggregate → squared sum → flips percentage when looking at sum

isolating the effects to make better predictions

② How do we know we have done a good job at classification?

→ Testing without cheating. Learning not memorizing

→ how well does the model perform on a dataset that it has never seen before

→ split up our data into training and seperate testing set.

→ use the training set to find patterns and create a model

→ use the testing set to evaluate the model on data, it has not seen before

how do we split data for training and testing minimize risk → ship as is.

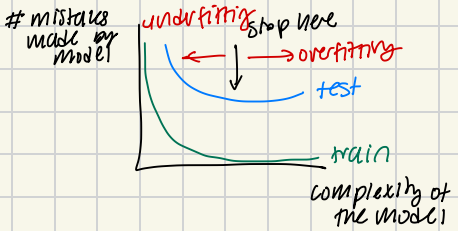
→ also allows us to check that we have not learned a model too specific to the model

→ overfitting vs. underfitting

→ goal is to capture general trends → watch out for outliers and noise points

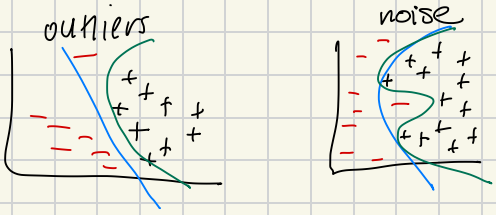
outliers are well defined, but how would we define noise points? depends if and how much we believe there is a relationship

looking at how the model performs on the testing set?



How do we avoid

Outliers versus noise → more noise / outliers we tried to account for → skew the data → overfitting but there is more to it



we want blue not green

Classification Recap

Training step:
→ create the model based on examples / data points in training set

Testing step:
→ use the model to fill in the blanks of testing set
→ compare the result of the model

instance-based classifier - just applying the model
→ finding matches → good for large datasets
→ but tricky for datasets with a lot of features and harder to find specific comb. of features
→ can choose two closest ones → but how do we choose from noise?
→ one is yes (malignant) and no (not malignant)

→ we want our model to learn patterns and not memorize the details of our dataset so it can generalize patterns

Underfitting vs. overfitting: we want to stop where our model is able to generalize well.

Ways of creating classifiers (How do we find a way to classify tumor type from age and tumor size?)

1.) Instance-Based Classifiers

Dataset is the model itself (applying the model)
→ use the stored

learning happens at prediction time, not training time because no training phase

→ there is no exact match, but similar examples that we can use to aggregate those to predict
1.) majority vote
2.) closest points — disadvantages: hard to set; varying levels
distance versus closest points → distance threshold → # of neighbors →

2.) K-Nearest Neighbor classifier:

Requires: Training set
Distance func
Value for k

How to classify unseen record
1.) compute distance of unseen record to all training records
2.) identify k nearest neighbors
3.) aggregate the labels of these k neighbors to predict the unseen record class (ex. majority rule)

different for weighted majority
→ we use the distance
↓ inversely proportional to the distance

aggregation methods: majority rule
weighted majority based on distance $w = \frac{1}{d^2}$