CS 412

Aiden Sirotkine

Spring 2025

Contents

Chapter 1

CS 412

Silly data mining summer class on Coursera lmao.

Extract patterns from vast swaths of data.

Data mining happens after data analysis where you classify and cluster and find patterns.

1.1 Types of Data

There's a whole bunch of types

1.1.1 Important Characteristics of Structured Data

- Dimensionality
- Sparsity
- Resolution

- Distribution

1.1.2 Nominal

Qualitative names of things. Categorical

1.1.3 Binary

2 possible values

There are symmetric and asymmetric binary attribute. Symmetic binary attributes are attributes where both classes are equally important, like biological sex. Asymmetric attributes are those where one of the attributes is more important than the other, like having a certain disease.

1.1.4 Ordinal

Ordered, ranked

1.1.5 Others

Just know there are more and they are self explanatory

1.1.6 Numeric

These are numbers

They can be either interval scaled where you have a range and they can be in that range, so there is no true zero point, or they can be ratio-scaled, where there is a true zero-point at 0.

There is also disrete numeric data which is discrete. They can also be continuous.

1.2 Central Tendency

Mean, median, mode, midrange.

There are weighted means. Means are also very sensitive to outliers unlike the median or mode.

The median splits the two halves of the data.

The mode is whatever datapoint has the most repeated values.

1.2.1 Median

If the median falls in the mth bin, meaning between m and $m + 1$, then a prediction for the median is

$$median \approx L_m + \frac{n/2 - F_{m-1}}{f_m} \times (L_{m+1} - L_m)$$

That's a predictor for the median.

1.2.2 Mode

$$mean - mode = 3 \times (mean - median)$$

The difference between the mean and the mode is 3 times the difference between the mean and the median. This formula holds for slightly skewed distributions.

Data can be skewed which means the mode is to the left or ride of the mean by a lot.

negatively skewed means a tail to the left (decreasing value)

Distributions can be multi-modal.

Symmetric data means

$$mean = median = mode$$

1.3 Dispersion Measures

Variance, Standard Deviation, Covariance, Correlation Coefficient.

You know the equations for the first 3

$$r = \frac{Cov(a, b)}{\sigma_a \sigma_b}$$

1.3.1 Normal Distribution

They're pretty cool

65-95-99.7 rule or whatever it is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

It also mimics binary random variables taken to infinity.

1.4 Statistical Tests

Chi-Squared Test tells you if a dataset follows a certain distribution.

You need a test statistic and a significance level.

You check if the p-value is above or below the significance level.

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

1.4.1 Null Hypothesis

There is no relationship between 2 variables

1.4.2 Degrees of Freedom

(num of row - 1) * (number of columns - 1)

or

number of categories - 1

so flipping a coin has 1 degree of freedom beccause 2-1 = 1

1.5 Contingency Tables

For a table where you're finding the expected value of males and fiction, You take the total number of males and the total number of fiction and you multiply them together and then divide by the total number of people in the entire table.

1.6 Data Visualization

It just makes it easier to read and easier to see correlations.

1.6.1 Quantile

Points taken at regularly separated intervals.

Percentile is a Quantile split 100 ways

Quartile is 25th, 50th, 75th percentile for 1st, 2nd, 3rd

You know box and whiskers plots

an outlier is 1.5 * IQR

Histograms plot data frequency in bins of certain ranges.

Bar charts look like histograms but they plot categorical data.

1.6.2 Quantile Plots

You sort the data and then plot it so the x-axis is the index or percentile and the y-axis is whatever you're measure

1.6.3 QQ Plot

Scatterplot where you sort two things and put 1 variable on 1 axis and the other variable on the other axis.

Very similar to scatter plots, but scatter plots have 2 different variables on each axis, while QQ plots have the same variable, but different samples for each datapoint.

1.6.4 Pixel Visualization

It just lets you see higher dimensional data.

1.6.5 Geometric Projections

You can have 3d scatterplots or matrices or landscapes and stuff

not as important as like QQ plots.

Chapter 2

Week 2

2.0.1 Similarity Measure

0 if different, 1 if the same

2.0.2 Dissimilarity/Proximity Measure

0 distance means the same.

Distance measure

You can compute the distances between different data matrices.

2.1 Categorical Data

Names, not numbers

2.1.1 Similarity

see how many match and use that percentage.

2.1.2 Contingency table

Make a 2 by 2 matrix and increment each box depending on the match when counting out the data.

The distance measure for symmetric binary data is.

$$\begin{bmatrix} q & r \\ s & t \end{bmatrix} \qquad \frac{r+t}{q+r+s+t}$$

The diagonals are important

For asymmetric data, you just ignore the unimportant variable (buying nothing)

Given a data matrix, you can make a dissimilarity matrix.

2.1.3 Minkowski Distance

$$d = \sqrt[p]{|x_{i1} - x_{j1}|^p + \dots}$$

A special case is the manhattan distance (take just the component vectors and sum them. Taxicab distance)

The supremum distance is the largest distance between any 2 individual dimensions.

2.1.4 Standardization

Make all data have a mean of 0 and a standard deviation σ

$$z = \frac{x - \mu}{\sigma}$$

2.1.5 Mean Absolute Deviation

$$S_f = \frac{1}{n}(|x_{1f} - m_f| + \ldots)$$

2.2 Ordinal Data

Standardize with

$$z = \frac{r_i - 1}{r_{last} - 1}$$

The -1 makes sure the thing starts at 0

You can also use a weighted average

2.2.1 Cosine Similarity

Angle between 2 vectors

$$sim(d_1, d_2) = \cos(\theta) = \frac{d_1 \cdot d_2}{|d_1| * |d_2|}$$

2.3 Probability Data

2.3.1 KL Divergence

The difference between 2 probability distributions over the same variable X

Measures the information lost when using q to approximate p

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln\left(\frac{p(x)}{q(x)}\right)$$

If q has a probability 0 for anything, just replace it with a very small ϵ and subtract ϵ/c from the c other probabilities.

2.3.2 Information Gain

Another way to think about it is the expected number of bits needed to sample from p starting at q

The amount of information gained from when one goes from prior probability q to posterior probability p

2.4 Data Cleaning

There are many tools to get rid of data discrepancies

2.4.1 Noisy Data

2.4.2 Binning

Sort data and put it into equal frequency bins and then turn each datapoint into the mean/median whatever of the bin.

2.4.3 Regression

Fit the datapoints along a curve

2.4.4 Clustering

Remove outliers

2.4.5 Semi-supervised

Use AI and a human to get rid of weird values.

2.4.6 Data Integration

Get the same data from multiple sources to try and get the most complete picture of the datapoints.

Clean the data with the mean/median/mode or the most recent data or something.

2.4.7 Redundancy

Don't count the same data multiple times.

2.5 Data Reduction

2.5.1 Linear Regression

Assume the data fits a model, find the parameters, get rid of the data.

2.5.2 Histograms, Clustering

You can also do a stratified sample which means you actively try to match the global probability distribution.

2.5.3 Data cube aggregation

You put all the data into a cube of nominal data and remove all unnecessary data.

2.5.4 Data compression

2.6 Data Transformation

Smoothing, Normalization, Discretization.

2.6.1 Min-Max Normalization

$$v' = \frac{v - min}{max - min}(newmax - newmin) + newmin$$

There's also z-score normalization

2.7 Dimensionality Reduction

When dimensions increase, the data becomes more and more sparse.

2^d possible combinations of d attributes

2.7.1 Supervised and Nonlinear

Feature selection and feature generation (make new features out of the existing data that might be useful)

2.7.2 Unsupervised and linear

PCA and feature extraction.

2.7.3 PCA

Take the k most important eigenvectors and eigenvalues and use just the most important eigenvectors to reconstruct/analyze the entire dataset.

Use the explained and cumulative variance to figure out how many eigenvectors to keep.

Chapter 3

Week 3

3.0.1 Supervised vs Unsupervised Learning

Supervised learning has labels already built in, while unsupervised learning clusters data based on patterns only.

3.0.2 Classification vs Regression

Classification guesses discrete or nominal labels.

Regression models continuous valued functions.

training set trains, validation set also trains, testing set is just for tests

3.1 Decision Trees

you can order data by just splitting points up based on their traits. Can handle all types of variables.

You do it recursively, splitting the data into the most meaningful groups.

NON-PARAMETRIC - no assumption of the data distribution.

unstable decision boundaries, and very sensitive to noise. Accuracy might not be perfect because of the simplicity. Perfect trees are NP hard and overfitting is common.

3.1.1 When to stop

all sorted, no attributes, or no datapoints

3.2 Splitting Measures

There's post pruning and pre-pruning and whatnot

there's information gain and entropy and stuff

3.2.1 Entropy

A measure of uncertaintly associated with a number

Entropy is always between 0 and 1

$$H(Y) = -\sum_{y \in Y} p_y \ln(p_y)$$

Higher entropy = higher uncertainty

You find the entropy for the whole system, and you change the decision tree based on if the entropy goes up or down for the whole system.

Y is the groups at each of the leaf nodes of the tree.

Information is the difference in the entropies before and after the split.

3.2.2 Conditional Entropy

You just look at the leaf nodes

$$H(Y|Patron) = \sum_{\text{leaf nodes after split}} p(x)H(Y|X = x)$$

You can have negative information gain but that means you do something stupid.

Information gain is biased towards attributes with a large number of values.

Use gain ratio instead

$$InformationGain = Entropy(pre) - Entropy(post)$$
$$GainRatio = \frac{Gain}{SplitInfo(A)}$$
$$SplitInfo(A) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \ln\left(\frac{|D_j|}{|D|}\right)$$

Whatever has the biggest gain ratio can be used for our splitting measure.

3.2.3 Gini Index (Impurity)

The Lower the Better

Used in binary trees.

For a dataset D with n classes

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2$$

Where p_j is the relative frequency of class j in D

For a split

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

And you just take the difference between the two.

Do some math and the Gini index is the probability of an error by random assignment.

Can be done with a multi-way split, but that wasnt its first use.

3.3 Comparisons

Information Gain: biased towards multivalued attributes

Gain Ratio: Tends to prefer unbalanced splits where 1 partition is much larger than the other

Gini Index: Biased to multi-valued attributes. Has difficulty when number of classes are large. favors equal sized partitions.

3.4 Lecture 10: Back to Decision Trees

3.4.1 Prepruning

Early stop. End when splits stop improving error

3.4.2 Post-pruning

Make the full minimized tree, then prune until cross-validation error is minimized.

Generally better then pre-pruning.

3.4.3 Random Forest

Make a whole bunch of decision trees and aggregate the predictions.

3.5 Baye's Theorem

Bayesian stats are based on belief because it usually is for events that cannot be repeated.

$$p(H|X) = \frac{p(X|H)p(H)}{p(X)} \propto p(X|H)p(H)$$

$p(X|H)$ is what we just observed, or the likelihood and $p(H)$ is the prior probability, and we are obtaining the posterior probability.

$p(X)$ is just the scaling factor.

Chapter 4

Week 4

4.1 Baye's Theorem with Multiple Hypotheses and Sequential Evidence

4.1.1 Sequential Evidence

$$p(H|X_1, X_2) = \frac{P(X_2|H, X_1)p(H|X_1)}{p(X_1, X_2)}$$

You just do bayes theorem twice I think

Assume X_1, X_2 are conditionally independent given H

$$p(H|X_1, X_2) = \frac{P(X_2|H)p(H|X_1)}{p(X_1, X_2)}$$

Then it becomes easier.

4.2 Naive Baye's Classifier

ASSUME FEATURES ARE CONDITIONALLY INDEPENDENT

Compute categorical features with frequency counts.

Continuous features likely use a Gaussian.

4.2.1 Sequential

You can use Baye's theorem to calculate the probability of an outcome given new information.

4.3 Linear Regression

Use math to map datapoints to a continuous formula.

Minimize the loss function to make a linear predictor

least-squares regression is the math used in lin alg. solves analytically.

4.4 Perceptron

Linear Classifier where you classify the top and the bottom of the classifier.

The line is adjusted if it misclassifies something

4.5 Logistic Regression

Uses a sigmoid function to classify between 2 classes.

If its below the sigmoid, its one class. Above is the other.

Returns a certain probability function that relates to the sigmoid.

$$\sigma = \frac{1}{1 + e^{-z}}$$

Minimize the negative log-likelihood.

4.5.1 Pros

Can handle many features. Fast, good. Robust, Interpretable.

Only works well if the decision boundary is linear.

4.6 Generative vs Discriminative Classifiers

Generative make a thing, Discriminative make a decision boundary and figure out everything else from there.

Chapter 5

Week 5

5.1 Model Evaluation

5.1.1 Confusion Matrix

It's a matrix of the number of actual traits in a dataset vs the number of predicted traits in a dataset, and it's used to examine how accurately a model models data.

true positive, false negative
false positive, true negative
Sensitivity = True Positive / Positive
Specificity = True Negative / Negative
Precision = True Positive / Predicted Positive
Recall = True Positive / Positive
Accuracy = True Positive + True Negative / All

5.1.2 F Measure

Gives weights to recall

$$F_s = \frac{(B^2+1)P \times R}{B^2P+R}$$

B = 1 means equal weight

In some cases precision might be more or less useful than recall.

5.1.3 ROC Curves

True positive rate TP/P on the y-axis and false positive rate FP/N on the bottom.

Receiver Operating Characcteristic.

It shows the true positive rate over the false positive rates, and as most models increase their true positive rate, their false positive rate also increases because they're just saying everything is true.

You can use the ROC curves of different classifiers to see what works best.

The area under the curve represents the quality of the model.

5.2 Classifier Evaluation

5.2.1 Holdout Method

Holdout Method is just training set and a validation set.

5.2.2 Cross-validation

Cross-Validation is the same as the holdout method but you do it multiple times over so that every part of the set becomes a validation set at some point.

You split it into a bunch

5.2.3 Leave-One-Out

self-explanatory.

5.2.4 Bootstrap Method

sample tuples with replacement.

The most common is the 0.632 bootstrap, where that much ends up in the training set and the rest is the validation set.

5.2.5 Parameters and Statistical Tests

Accuracy, Speed, Robustness, Scalability, Interpretability.

You can do a t-test to compare models.

5.3 Lazy Learning

Stores training data and waits until given a test tuple.

less training time, more predicting time.

example: Nearest Neighbor.

You need proper indexing/decent algorithms because lazy learning can be computationally expensive.

Must commit to a single hypothesis for the whole dataset.

5.3.1 KNN

Look at the nearest points and do a majority vote and boom youve classified it.

5.4 Ensemble Methods

5.5 Bagging and Random Forest

EACH TREE HAS RANDOMLY SELECTED FEATURES.

Bagging is training a bajillion different models with different training sets and put em all together.

boot + agg(regating models) = bagging

Random forest is bagging + a decision tree. You have different features for some models and they're selected randomly.

The different models have different features.

5.6 Boosting

Sequentially put a datapoint through classifiers with each one mending the mistakes of the last.

5.6.1 Adaboost

Adaptive boosting.

Assign initial weights to a dataset, fix the incorrect classifications by changing the weights. Repeat until classifier is good.

5.6.2 Gradient Boosting

Incrementally add procedurally weaker classifiers to the model to decrease the loss function.

You need a differentiable loss function.

5.6.3 Regression Tree

decision tree but for continuous variables.

Chapter 6

Week 6

6.1 Class Imbalance

Alot of methods assume both classes have equal error cost. This however is not true for many real-life samples (rare disease diagnosis, credit card fraud.)

You can oversample from the minority and undersample from the majority.

6.1.1 Threshold Moving

Increase the chance of classification for the minority to decrease the chance of the false negatives (Better to be safe than sorry).

6.1.2 Class Weight Adjusting

Make false negatives more penalizing.

6.2 Bayesian Belief Network

Given a di-graph of dependent probabilities, find the total probability of an outcome given the data.

6.2.1 Training

If the structure is known, and the variables are observable, compute the Conditional Probability Table entries, and estimate the probabilities analytically.

If the structure is known, but the variables are hidden, use the training data to predict the unknown probabilities/parameters. Use gradient ascent to maximize probabilities.

6.2.2 Plate Notation

You index graphs that are all related and put them all into 1 square.

6.3 Support Vector Machines

You use a non-linear mapping to map your data into a higher dimension such that you only need a linear hyperplane to separate your classes.

You can also do the mapping using a kernel function.

6.3.1 IF-THEN Rules

Seeing what previous traits correlation to other classifications.

6.3.2 Size Ordering

Assign highest priority to triggering rules that have the ”toughest” requirement (with the most attribute tests)

6.3.3 Class-Based Ordering

Decreasing order of prevalence or misclassification cost per class.

6.3.4 Rule-based Ordering

Rules are organized into one long priority list according to some measure or expert opinion.

6.3.5 Rule Induction

Rules are mutually exclusive and exhaustive which means no two rules are triggered by the same tuple, and all combinations of attributes have a rule.

Start with a list of 0 rules.
Add a rule
Remove the point classified by the rule.
repeat until terminating condition.

6.4 Pattern-Based Classification

Learn patterns

6.5 Semi-Supervised Learning

Use labeled AND unlabeled data to train a classifier.

6.5.1 Self-Training

Train with the labeled data. Predict the unlabeled data. The most confident predictions get added to the labeled data. Repeat.

6.5.2 Co-Training

Use two features with mutually independently features. Give the most confident predictions to THE OTHER List to not reinforce errors.

6.5.3 Active Learning

The learner queries a human expert to have only the most informative data for training. It specifically queries the least confident data in the set.

You can also use a learning curve

6.5.4 Transfer Learning

Use a separate classifier to help train with unlabeled data.

6.5.5 Weak Supervision

Use noisier less high quality data to train the data instead of our fancy labeled training data.

You just ask people or use wikipedia for your weak data.