DATA MINING NOTES

INTRODUCTION

BAYESIANBRAT*

Data Mining:

→ Data mining is the process of finding useful patterns, trends and relationship in large datasets. It goes beyond traditional statistics because data sizes are much bigger, complex and often less structured.

1. Not all traditional statistical approaches are useful

In traditional statistics, techniques like

→ p-values
→ hypothesis testing
→ t-tests
→ Chi-square tests are widely used.

→ But when you have huge datasets ($n = 100,000$ or more), these tests almost detect "statistical significance" – even for tiny, ~~even~~ unimportant differences.

Example: 100,000 data points, even a 0.001 difference in average income between two groups may appear significant it might not be practically meaningful

2. Allows new methods:

More expressive models:
Instead of simple linear regression, we use decision trees, random forests, neural networks etc to find and capture complex patterns.

Relax assumptions:
Traditional statistics often assume things like "data is normally distributed" or variables are independent".

Data mining models don't always need those assumptions, they can work directly with raw-messy real world data.

Empirical model building:

In data mining, we test models directly on data instead of just proving them with math.
proving

Direct measures of model quality:

Instead of relying on p-values, we measure things like accuracy precision, recall, F1 score, RMSE, ROC-AUC etc.

These tells us how good the model is at prediction and classification.

* Data Mining textbook definition:

The process of extracting valid, previously unknown comprehensible information from large database in order to improve and optimize business decisions.

→ Categories of Tasks

Is response / target attribute available?

- Yes → Supervised Learning
  Target numerical: Regression. Weight, revenue, glucose, time, velocity.
  Target categorical: Classification:
  Disease present, purchase, defect present, object type, sound type, document type.
- No → No Supervised Learning
  - Cluster analysis
    Group documents, images, tweets, segment customers, process campaigns.

Semi supervised learning:

Semi-supervised learning is a hybrid approach that uses small amount of labeled data and a large scale amount of unlabeled data to train a model.

Example:
You have 1000 handwritten digits, but only 100 are labeled (0-9).
The algorithm First learns patterns from all images (unsupervised part), then fine-tunes using the 100 labeled examples (supervised part).

Result → good accuracy with much less "labeled data.

Algorithm Techniques:

- Label propogation / Label spreading.
- Self-training (train, label the unlabled, retrain)
- Graph-based learning
- Semi supervised neural networks

Self - Supervised learning :

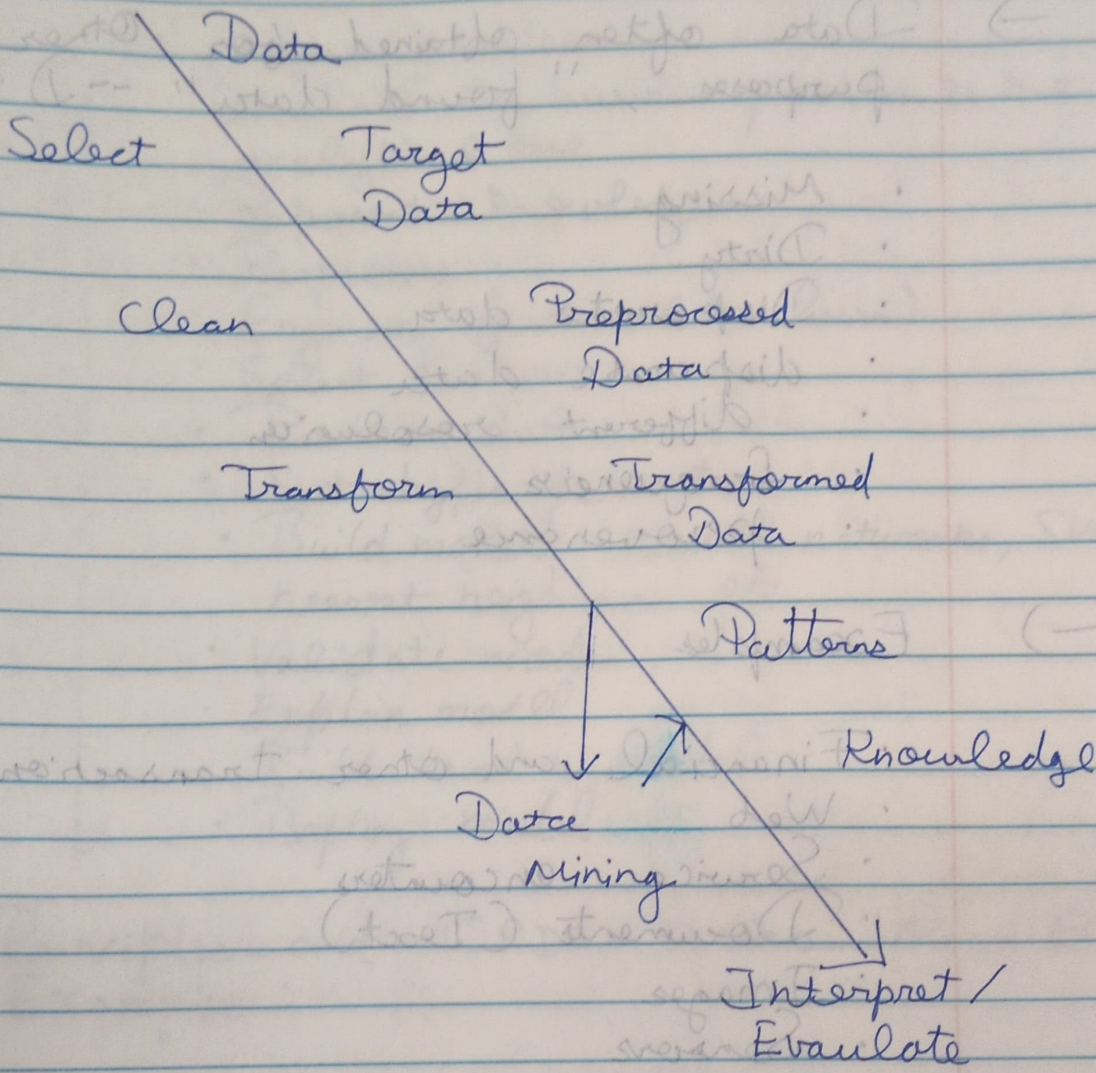
Self supervised learning is a special type of unsupervised learning where the system creates its own labels from the data itself.

It's still trained like a supervised model ; but the "supervision" comes from the data's internal structure and not human labeling.

→ Example

a. Text (e.g. GPT models):
→ Take a sentence : "The cat sat on the ____."
→ Mask one word ("mat"), and train the model to predict it.
→ The missing word is self-generated label.

Knowledge Discovery Process



Characteristics of Data:

→ Tremendous resources allocated to databases.

→ Data often obtained for other purposes, "found data" -- D Brain

- Missing
- Dirty
- Disparate data
- disparate data
- different resolutions
- frequencies
- provenance

→ Examples

- Financial and other transactions
- Web
- Service encounters
- Documents (Text)
- Images
- Sensors
- Communications
- Genomics

Processes and Models

Top items have large effect on accuracy.

- Define problem
- Translate the business problem to the analytical task.
- Represent knowledge
- Preprocess
- Engineer features (or learn)
- Select features
- Partition data
- Set hyper-parameters
- Build model: neural networks, SVM, nearest neighbours etc.
- Validate model
- Explain model
- Build trust in model
- Deploy model

★ Principles and CRICA 2000 "New Economy" Examples:

- Data analysis principles still apply.

→ Curse of dimensionality:
As datasets have more and more features, analysis become harder.

→ This is fundamental principle and it doesn't go away because we have big data.

→ Ongoing battle between "bad" and "evil": This refers to the challenge of distinguishing bad models (simple mistakes) vs evil models (overfitted).

In data mining, there's always danger of overfitting.

→ Trendy models today likely to be replaced in a few years, but concepts still persist.

→ In short: Algorithms like decision trees, SVM, deeplearning etc come and go in popularity.

→ But the underlying principles remain the same.

Curse of dimensionality:

→ The curse of dimensionality is a concept that explains why analyzing ad organizing data becomes much harder as number of features increases.

→ Imagine each feature is a new column in a spreadsheet.

→ Think of a single dimension problem, like data with just one feature, say a person's height. You can easily see how data is distributed.

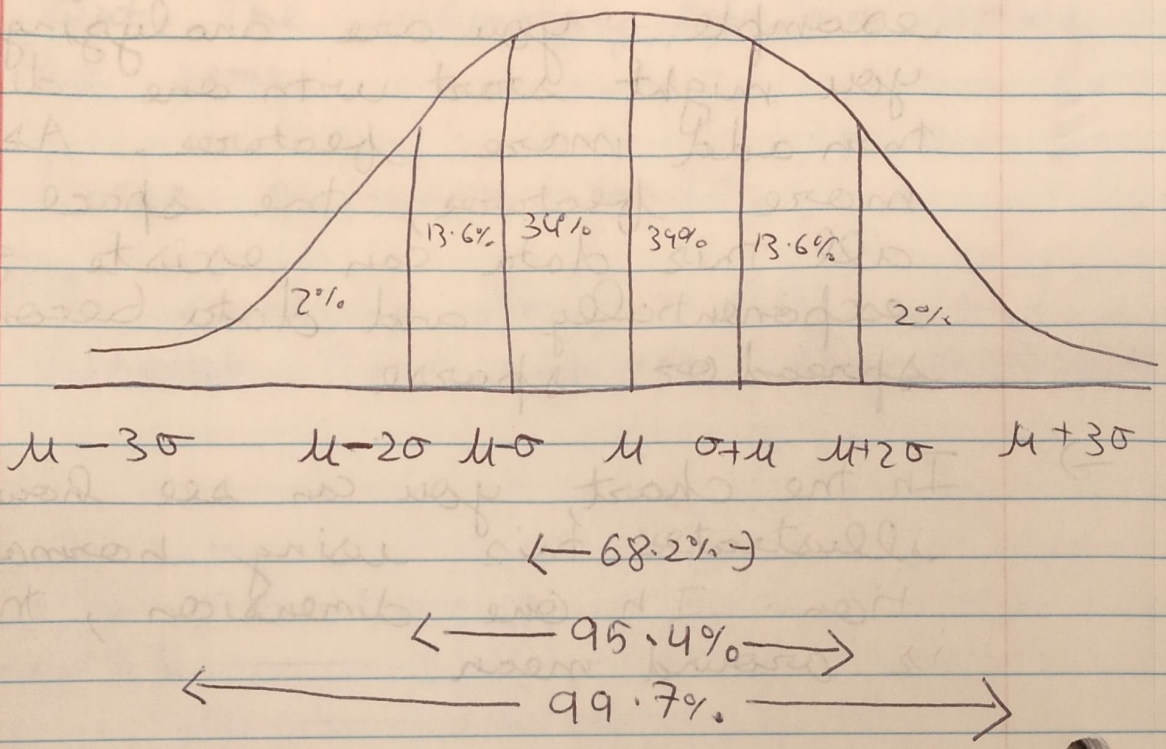
→ Let's add some more features, example, you are analyzing houses, you might start with one dimension then add more features. As you add more features, the space where all this data can exists grows exponentially and data become very spread or sparse.

→ In the chart, you can see how it illustrates this using normal distribution. In one dimension, the region is around mean

- Assume independent variables
- One dimension:
  - Normal curve region:
    mean +/- one sigma contains 68% of the data.

- When you have 10 dimensions, the "same" region only contains $0.68^{10} = 2\%$ of the data. That's a huge drop.

- With 50 dimensions, the same region contains incraditably small fraction of data.



→ As the data is sparse, probability of finding it gets much lower.

★ Example to train a linear model:

- Linear model
- 2000 instances
- 19 attributes (predictors)
- 20 parameters
- ignore data between and within subjects

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{19} x_{19}$$

Linear model estimates values for β's to minimize loss of Sum of squared errors.

$$SSE = \sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2$$

$$SST = \sum_{i=0}^{N-1} (y_i - \bar{y})^2$$

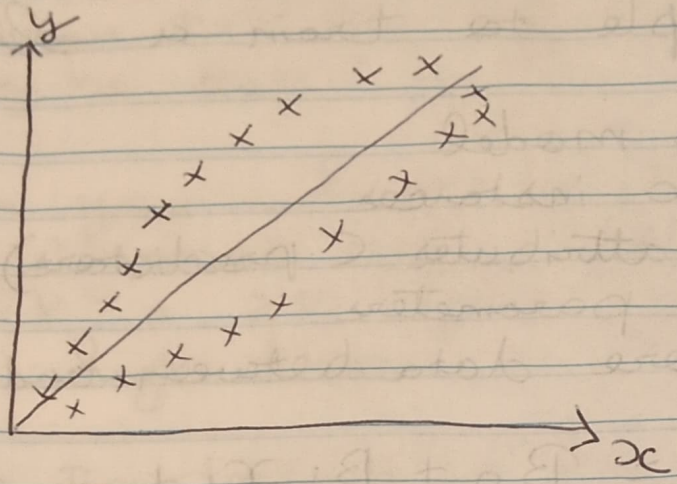
$$R^2 = 1 - \frac{SSE}{SST}$$

→ R^2 is used to measure model quality.
→ R^2 is interpreted as proportion of variance in y explained by model.
→ Greater R^2 value is desirable

Polynomial linear regression:

Lets see three cases:

1.



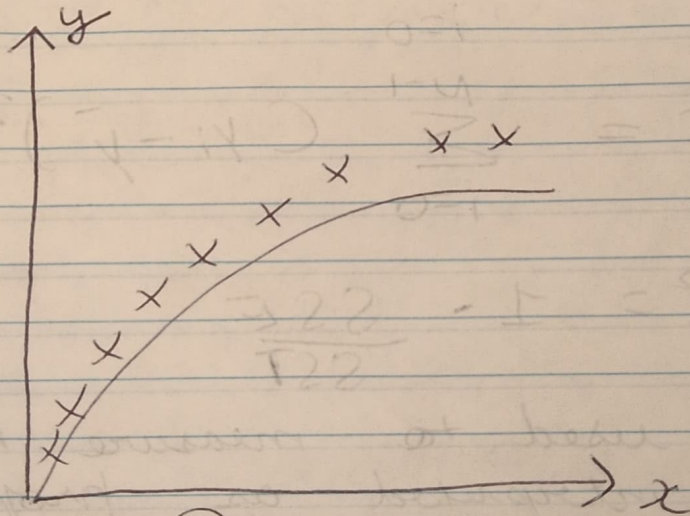
Degree = 1

High Bias and High Variance

Training data accuracy is low and test data cacuracy is Low too

This is the case of underfitting

2.

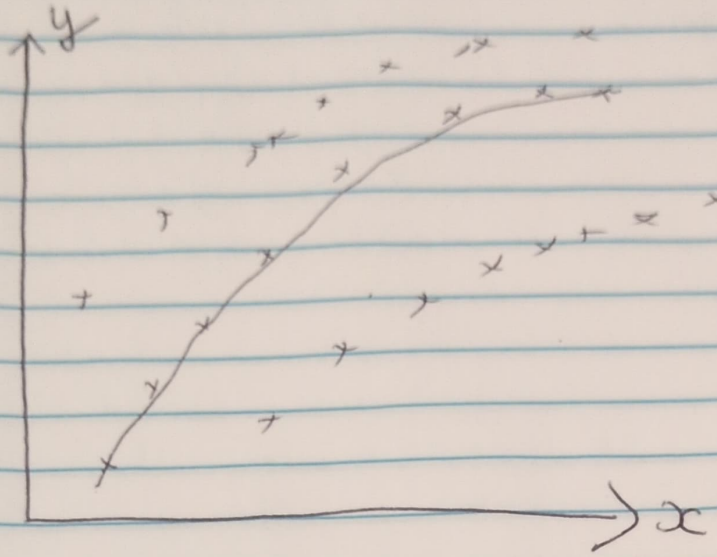


Degree = 2

Low Bias and Low Variance

Training data acc ↑ Test data acc ↑

3.



Degree = 4

Low Bias ~~Low~~ High Variance

Training data accuracy is high
Test data accuracy is low

This is the case of Overfitting