Data Processing Essentials for Building Predictive Models with Python (Data Quality Series)

General Poll

How comfortable are you programming in Python?

- Very comfortable
- Somewhat comfortable
- Never programmed in Python before

General Poll

How familiar are you with machine learning?

- 3+ years experience
- 1+ years experience
- Never worked with machine learning models before

General Poll

Have you attend Data Cleaning Essentials (Part - I) and Data Preparation Essentials (Part - II) of the data quality series?

- Yes
- No, but I am familiar with the topics covered (cleaning data, standardization, working with text data)

Problems with Data

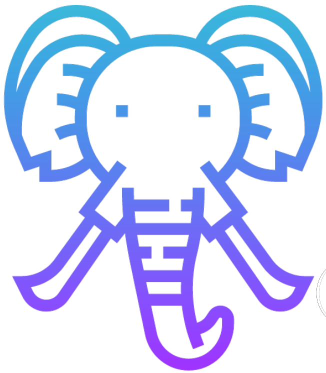
Insufficient data	Too much data	Non-representative data
Missing data	Duplicate data	Outliers

Problems with Data

Insufficient data	Too much data	Non-representative data
Missing data	Duplicate data	Outliers

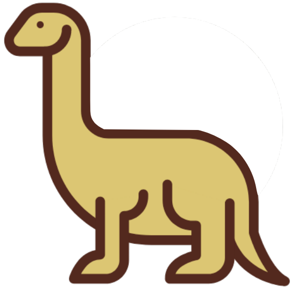
Too Much Data

- Data might be excessive in two ways
    - Curse of dimensionality: Too many columns
    - Outdated historical data: Too many rows



Outdated Data

- Outdated historical data is a serious issue in specific applications
  - Financial trading
- Usually requires human expert to judge which rows to leave out



Curse of Dimensionality

- Two specific problems arise when too much data is available
  - Deciding which data is actually relevant
  - Aggregating very low-level data into useful features

Curse of Dimensionality: As number of **x** variables grows, working with data poses several problems

Curse of Dimensionality

Visualizing data

Training ML models

Using ML models for prediction

Problems in Visualization

- Exploratory Data Analysis (EDA) is an essential precursor to model building
- Essential for
  - Identifying outliers
  - Detecting anomalies
  - Choosing functional form of relationships
- Higher dimensional data may not be explored properly before fitting predictive models

Problems in Training

- Training is the process of finding best model parameters
- Complex models have thousands of parameter values
- If we do not train for long enough, model parameters may not have converged to best possible values

Problems in Training

- Number of parameters to be found grows rapidly with dimensionality
- Extremely time-consuming, may require additional resources
- Training on the cloud can get very expensive

Problems in Prediction

- As dimensionality grows, size of search space explodes
- Higher number of feature leads to data sparsity
- Higher risk of overfitting on the training data

Curse of Dimensionality

- Easier problems to solve
  - Feature selection: Deciding which data is actually relevant
  - Feature engineering: Aggregating very low-level data into useful features
  - Dimensionality Reduction: Reduce complexity without losing information

Drawbacks of Reducing Complexity

Information loss

Performance degradation

Computational intensive

Transformed features hard to interpret

Feature Selection and Dimensionality Reduction

Reducing Complexity

Feature Selection

Filter methods

Wrapper methods

Embedded methods

Dimensionality Reduction

Projection

Manifold Learning

Autoencoding

Reducing Complexity

Feature Selection

Dimensionality Reduction

Choose the most relevant X variables from the existing data

Reducing Complexity

Feature Selection

Dimensionality Reduction

Transform the original X variables into new dimensions

Reducing Complexity

- Feature Selection
  - Filter methods
  - Wrapper methods
  - Embedded methods
- Dimensionality Reduction
  - Projection
  - Manifold Learning
  - Autoencoding

Feature Selection

Choosing Feature Selection

- Many X-variables i.e. many features in the input data
- Most of which contain little information
- Several features might be irrelevant
- Some of which are very meaningful
- Meaningful variables are independent of each other

Filter Methods

- Features selected based on statistical properties of features
- Either individually (univariate) or jointly (multi-variate)

Statistical Techniques

Variance Thresholding

Chi-square Test

ANOVA

Variance Thresholding

If all points have the same value for an X-variable, that variable adds no information.

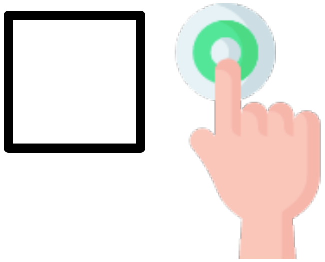
Extend this idea and drop columns with variance below a minimum threshold.

Chi-square Feature Selection

For each X-variable, use the Chi-square test to evaluate whether that variable and Y are independent. If yes, drop that feature. Used for categorical X and Y.

Chi-square Feature Selection

- Does observed data deviate from those expected in a particular analysis?
- Tests the effect of one variable on the target, univariate analysis
- Sum of the squared difference between observed and expected data in all categories



ANOVA

ANalysis ***O***f ***VA***riance

ANOVA

Looks across multiple groups of populations, compares their means to produce one score and one significance value

ANOVA Feature Selection

For each X-variable, use the ANOVA F-test to check whether mean of Y category varies for each distinct value of X. If not, drop that X-variable.

Wrapper Methods

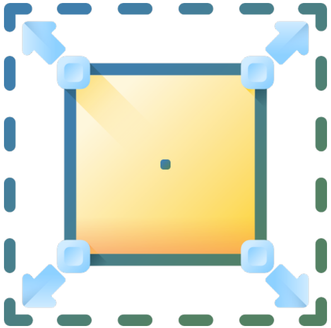
- Features are chosen by building different candidate models
- Forward and backward stepwise regression are examples

Wrapper Methods

- Each candidate model has different subset of features
- However all candidate models are similar in structure
- Features may be added or dropped to see whether the model improves

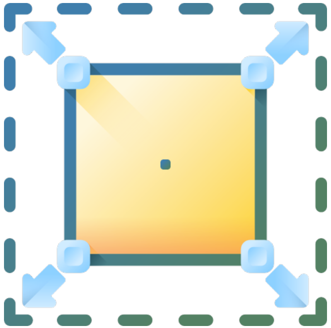
Embedded Methods

- Features (columns) selected during model training
- Feature selection effectively embedded within modeling
- Only specific types of models perform feature selection



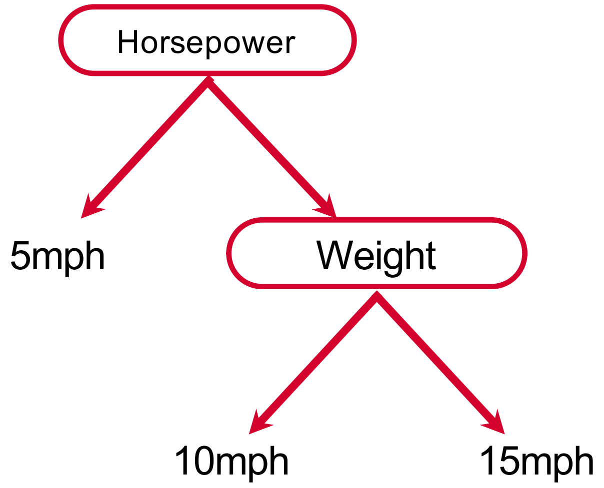
Embedded Feature Selection

- Some machine learning algorithms automatically perform feature selection
  - Decision trees
  - Lasso regression



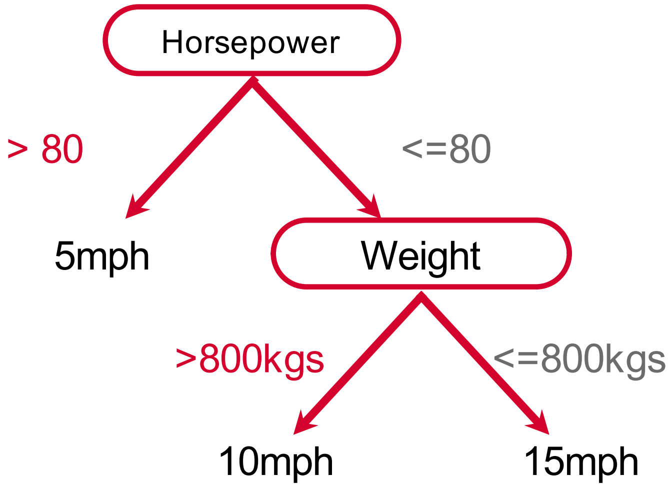
Decision trees set up a tree structure on training data which helps make **decisions** based on **rules**

Decisions Based on Rules



Predict car mileage based on car attributes

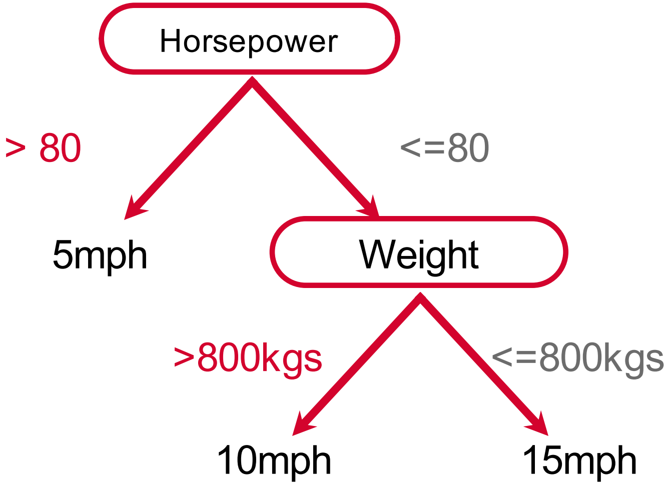
Decisions Based on Rules



Every decision node is based on a threshold

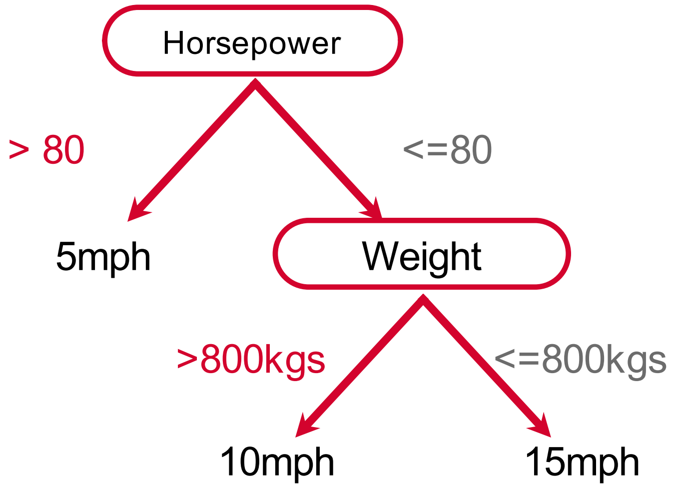
Decision Tree

- Order of decision variables matters
- Rules and order found using ML



Decision Tree

- Order of decision variables matters
- Order determines feature importance



Lasso regression performs model selection by setting coefficients of unimportant features to be close to zero

Ordinary MSE Regression

Minimize $\sqrt{(y^{actual} - y^{predicted})^2}$

To find

A, B

The value of A and B define the “best fit” line

$$y = A + Bx$$

Lasso Regression

Minimize $\sqrt{(y^{actual} - y^{predicted})^2}$ $+ \alpha (|A| + |B|)$

To find

A, B

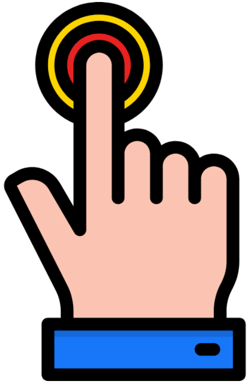
α is a hyperparameter

The value of A and B still define the “best fit” line

$$y = A + Bx$$

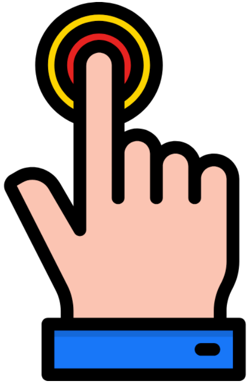
Lasso Regression

- Add penalty for large coefficients
- Penalty term is L-1 norm of coefficients
- Penalty weighted by hyperparameter α



Lasso Regression

- “Lasso” ~ Least Absolute Shrinkage and Selection Operator
- Math is complex
- No closed form, needs numeric solution
- Performs model selection by setting the coefficients of unimportant features to zero



Hands-on demos on feature selection

Polls

Poll 1

Which of the following is an example of a feature selection technique?

- Chi2
- Principal Components Analysis
- Manifold Learning
- Factor Analysis

Poll 1

Which of the following is an example of a feature selection technique?

- **Chi2**
- Principal Components Analysis
- Manifold Learning
- Factor Analysis

Poll 2

Which of the following statistical tests could you use to find features that are the most statistically significant?

- Recursive Feature Elimination
- ANOVA
- PCA

Poll 2

Which of the following statistical tests could you use to find features that are the most statistically significant?

- Recursive Feature Elimination
- **ANOVA**
- PCA

Poll 3

Which of the following is NOT a feature selection technique?

- Wrapper methods
- Filter methods
- Projection methods

Poll 3

Which of the following is NOT a feature selection technique?

- Wrapper methods
- Filter methods
- **Projection methods**

Poll 4

Which of the following regression algorithms sets the coefficients of unimportant features to zero?

- Support vectors
- Ridge
- Ordinary least squares

Poll 4

Which of the following regression algorithms sets the coefficients of unimportant features to zero?

- Support vectors
- Ridge
- Ordinary least squares

Dimensionality Reduction

Reducing Complexity

Feature Selection

Dimensionality Reduction

Projection

Autoencoding

Manifold Learning

Project the data onto new axes, entirely re-orients the data e.g. PCA, factor analysis, LDA

Reducing Complexity

Feature Selection

Dimensionality Reduction

Projection

Autoencoding

Manifold Learning

Unrolls the data, onto lower dimensional space e.g. MDS, LLE, Isomap

Reducing Complexity

Feature Selection

Dimensionality Reduction

Projection

Autoencoding

Manifold Learning

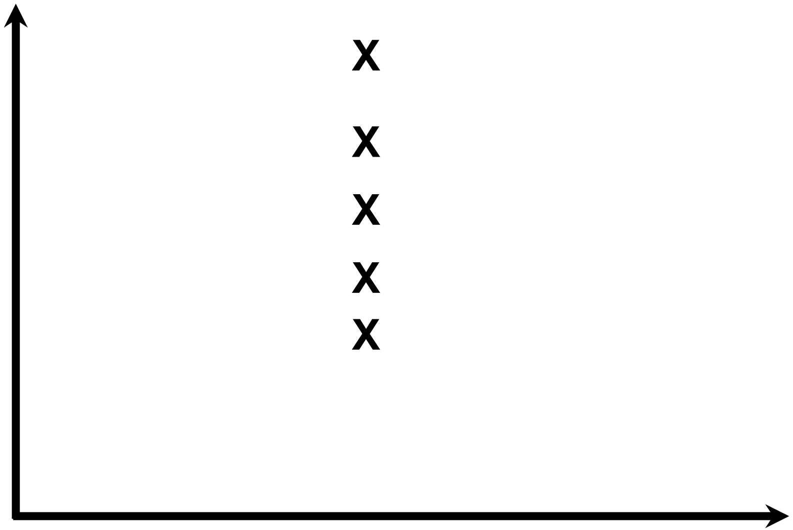
Extract efficient representations
i.e. latent features of complex data

Choosing Projection Techniques

- Large number of X-variables in the input data
- Most of which are meaningful
- Highly correlated to each other i.e. linearly related to each other
- Reduce multicollinearity in data

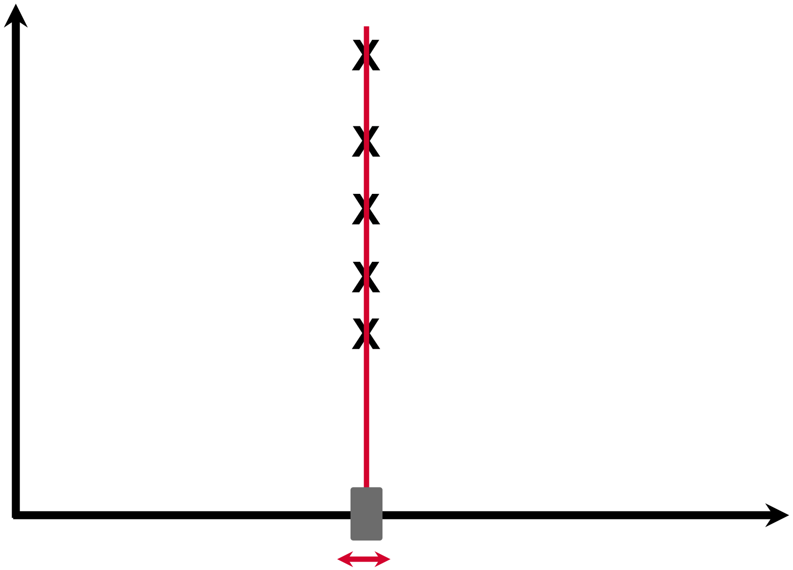
- Principal Component Analysis (PCA) and Factor Analysis for regression models
- Linear Discriminant Analysis (LDA) for classification models

Representing Data



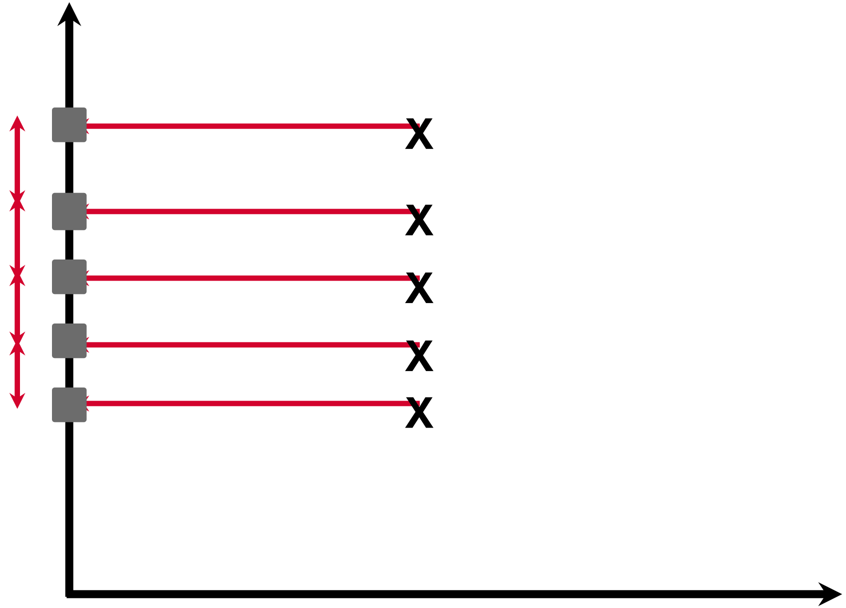
How many dimensions do we need to represent this data?

Choosing Dimensions?



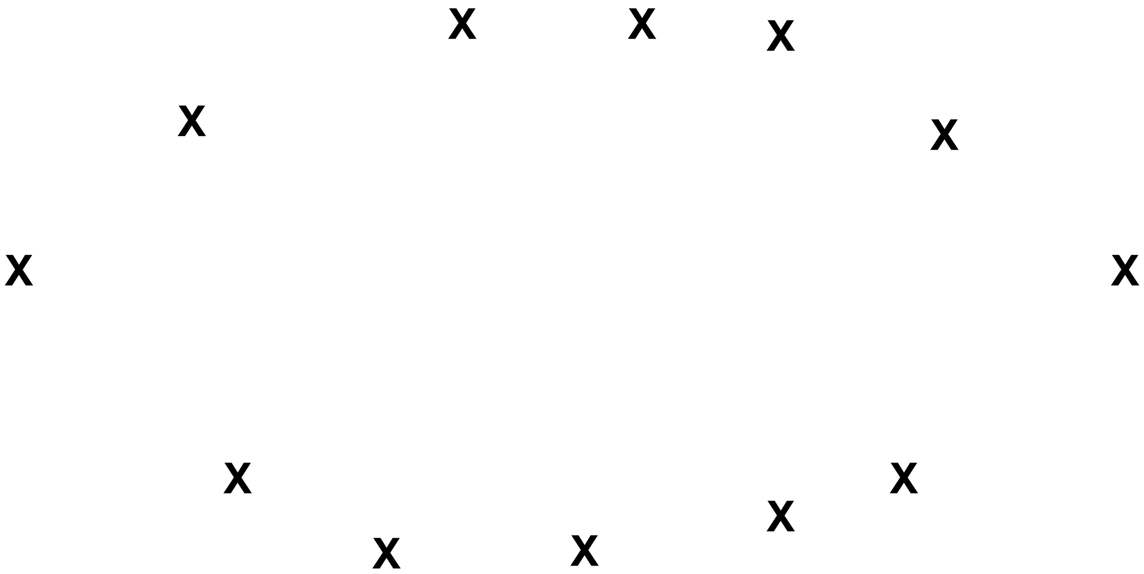
Is this a good choice?

Choosing Dimensions?



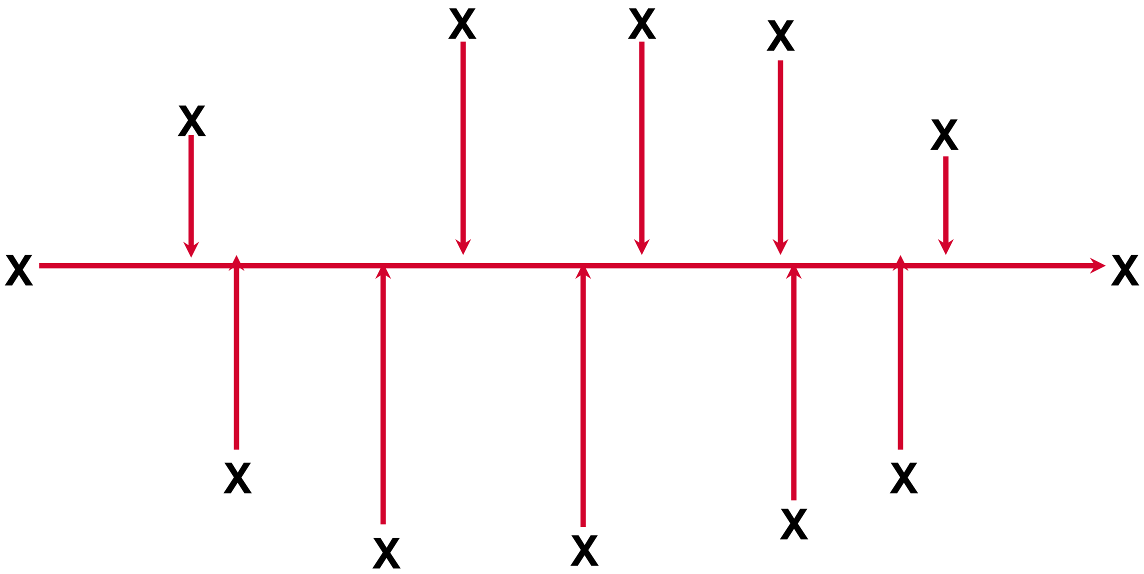
What about this?

Understanding PCA

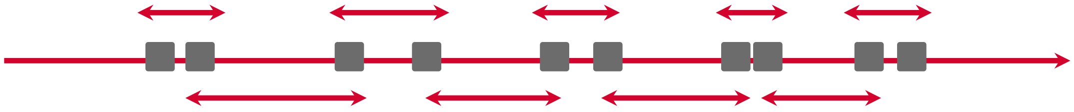


Find the “best” directions to represent this data

Understanding PCA

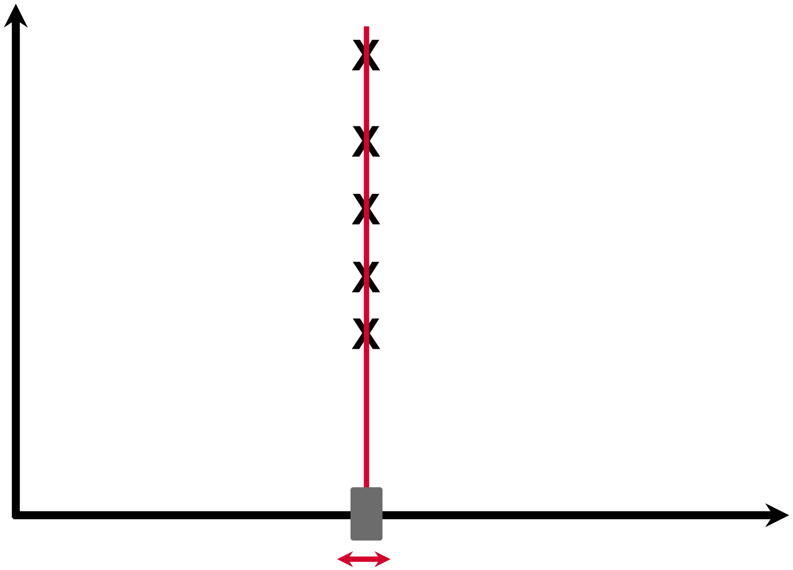


Understanding PCA



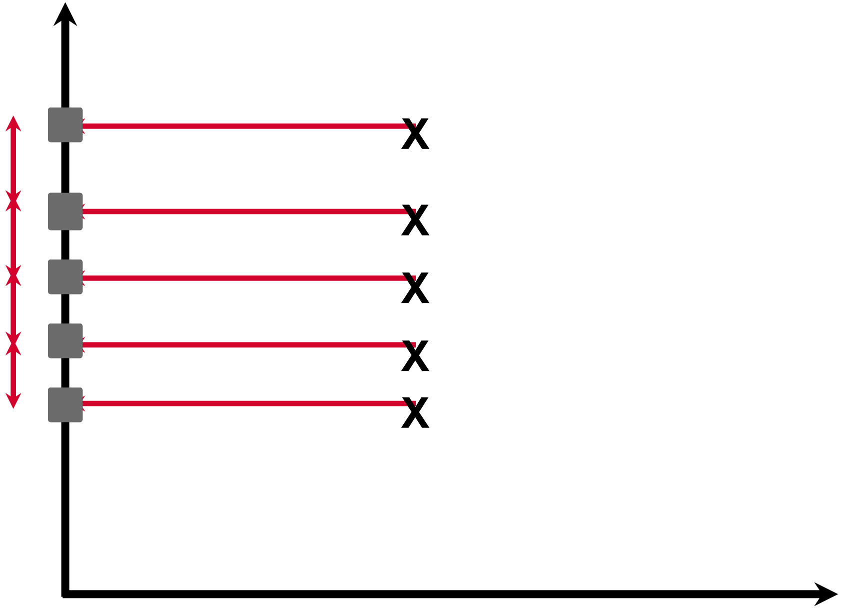
Try maximizing the distance between data points on the projected axis

Bad Projection



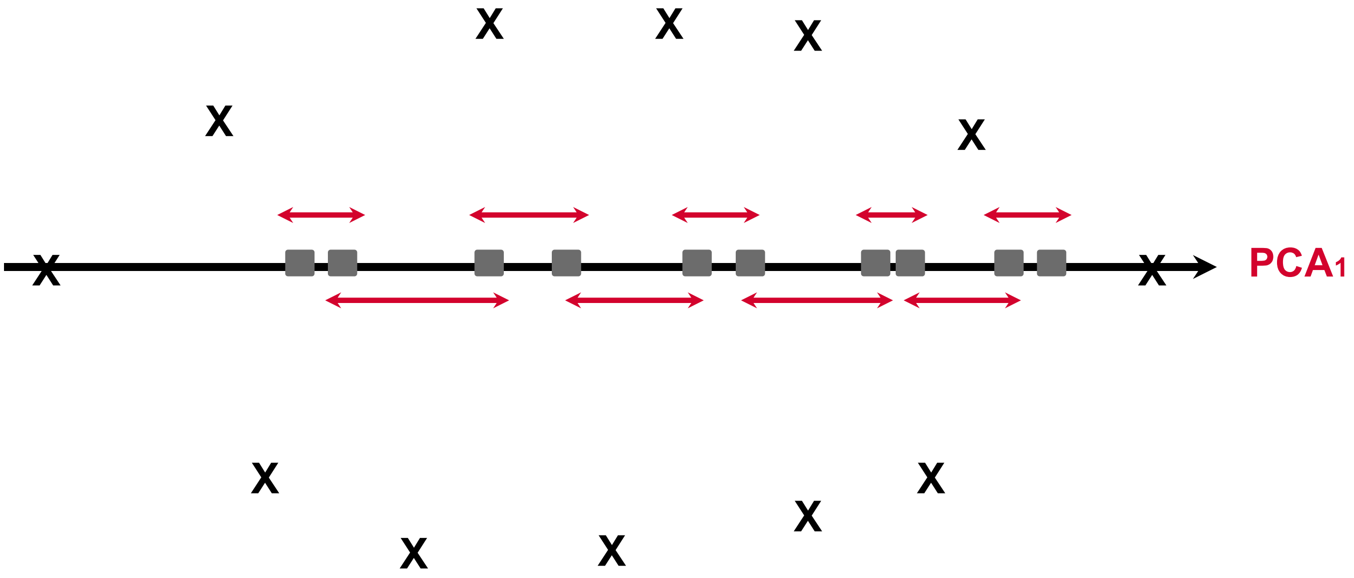
Information is lost here

Good Projection



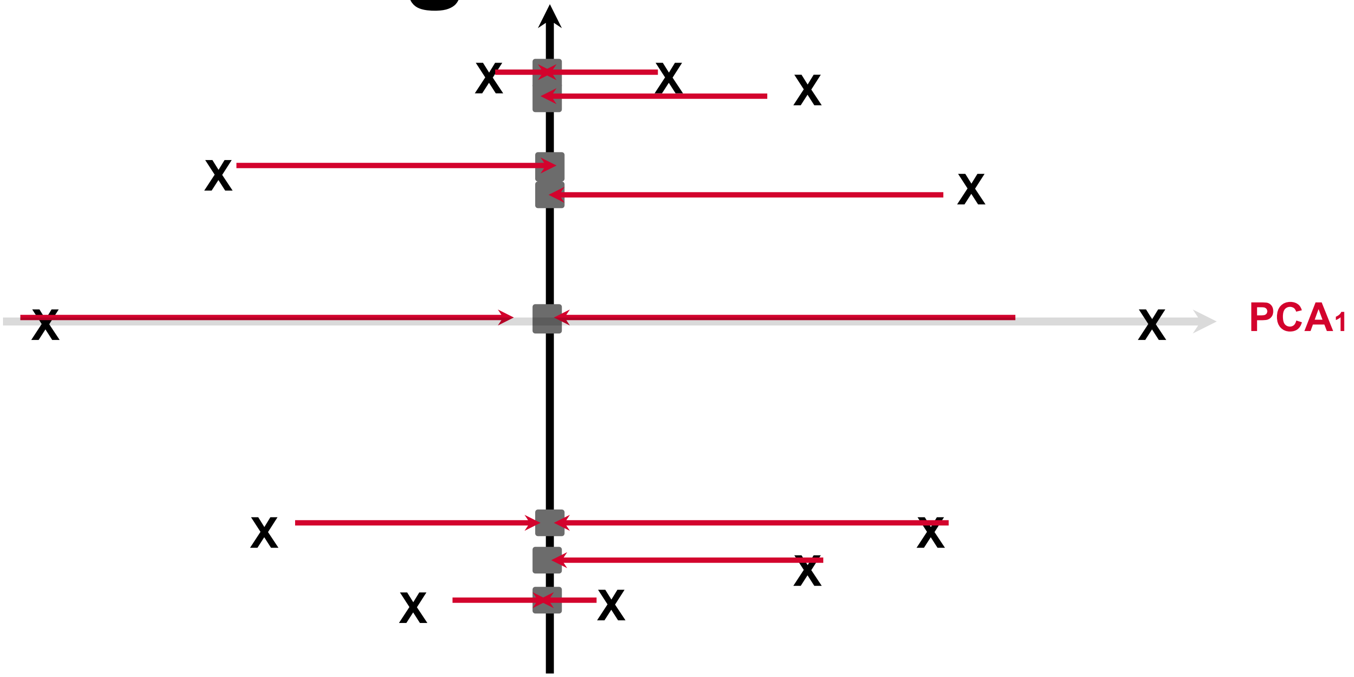
Information is preserved

Understanding PCA



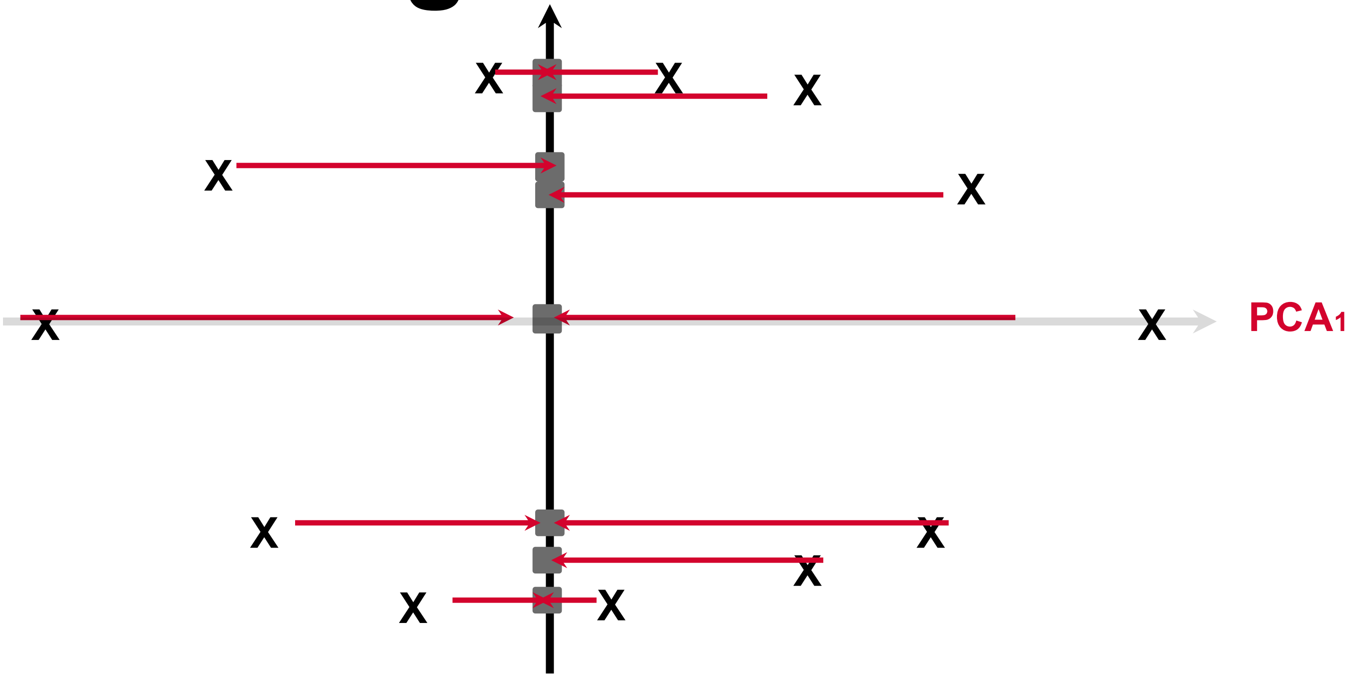
The direction along which this variance is maximized
is the first principal component of the original data

Understanding PCA



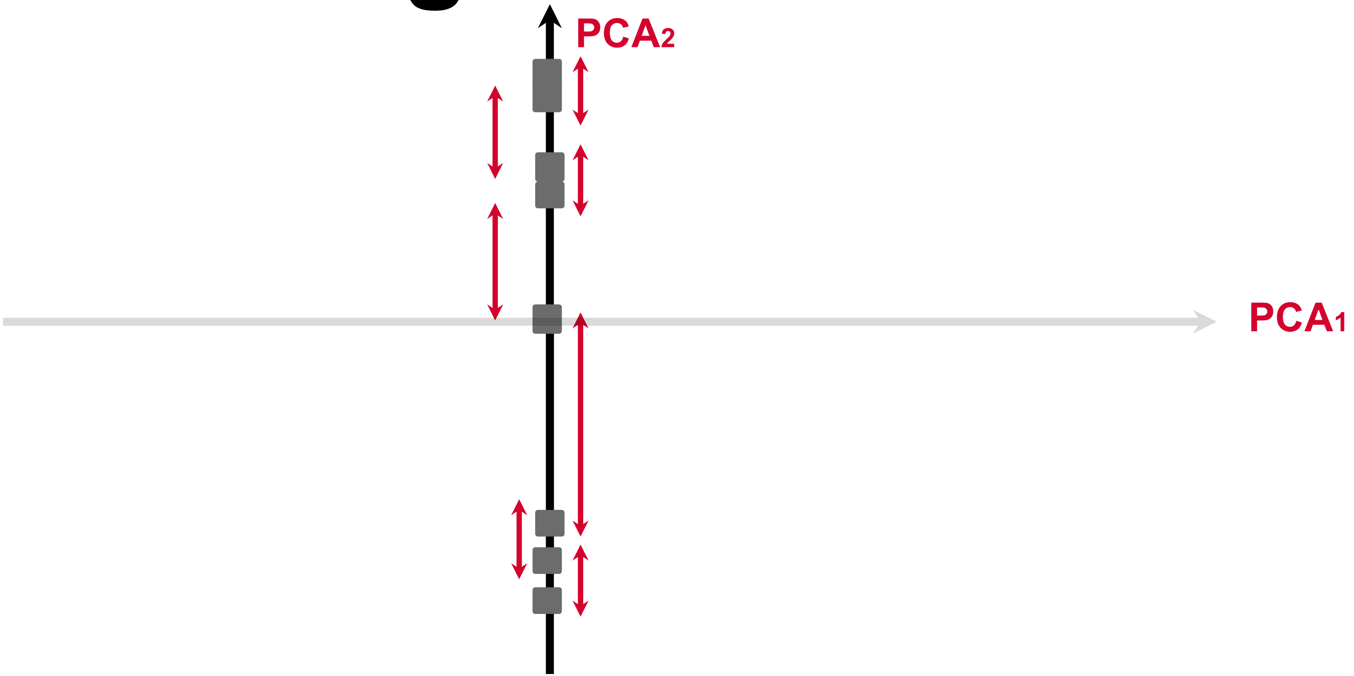
The second principal component, is at right angles to the first

Understanding PCA



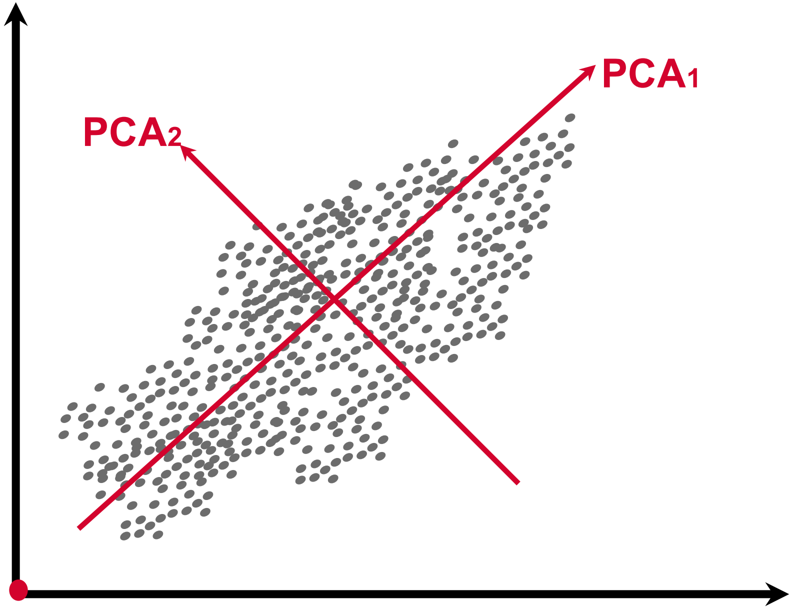
Directions at right angles help express the most variation with the smallest number of directions

Understanding PCA



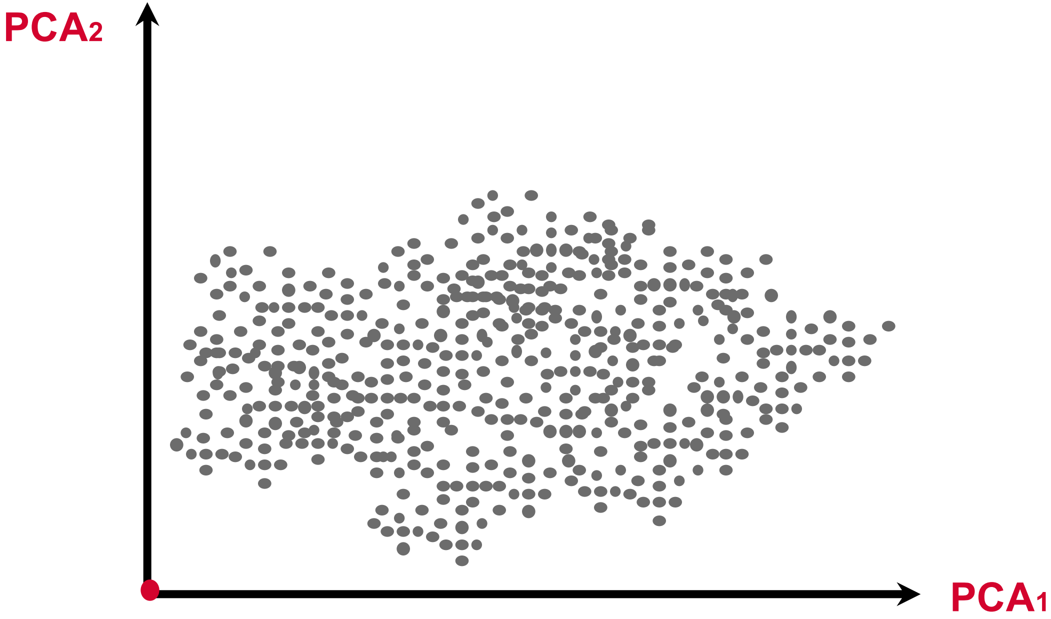
The variances are clearly smaller along this
second principal component than along the first

Understanding PCA



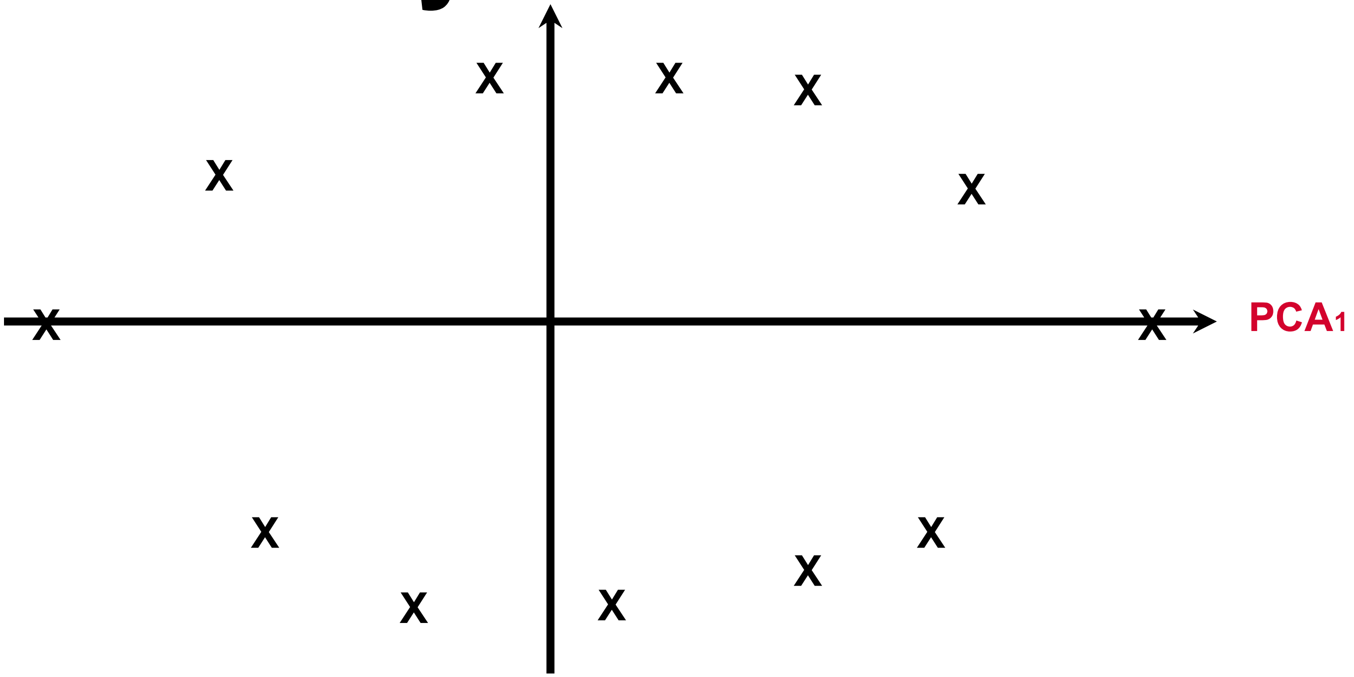
In general, there are as many principal components as there are dimensions in the original data

Understanding PCA



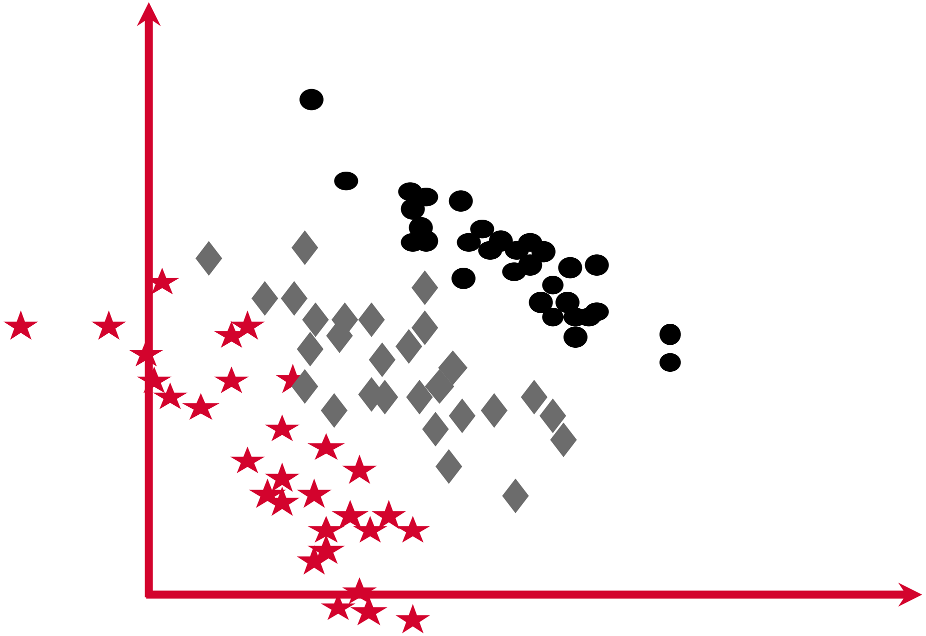
Re-orient the data along these new axes

Dimensionality Reduction

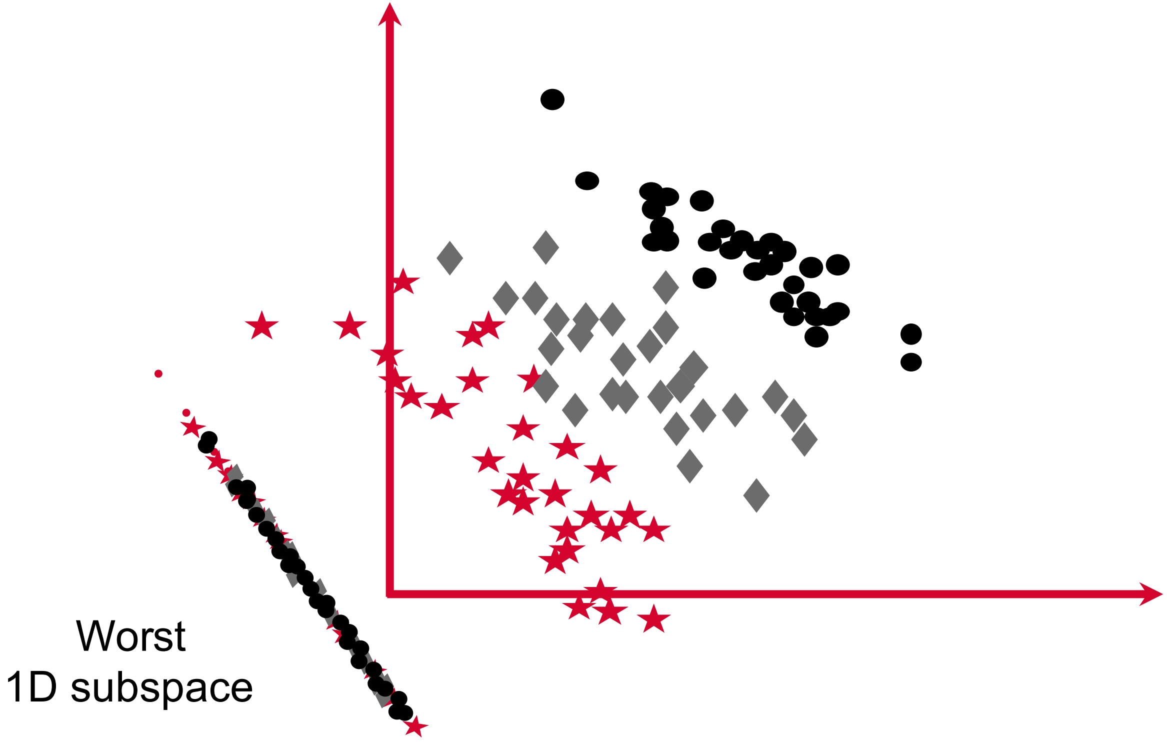


If the variance along the second principal component is small enough, we can just ignore it and use just 1 dimension to represent the data

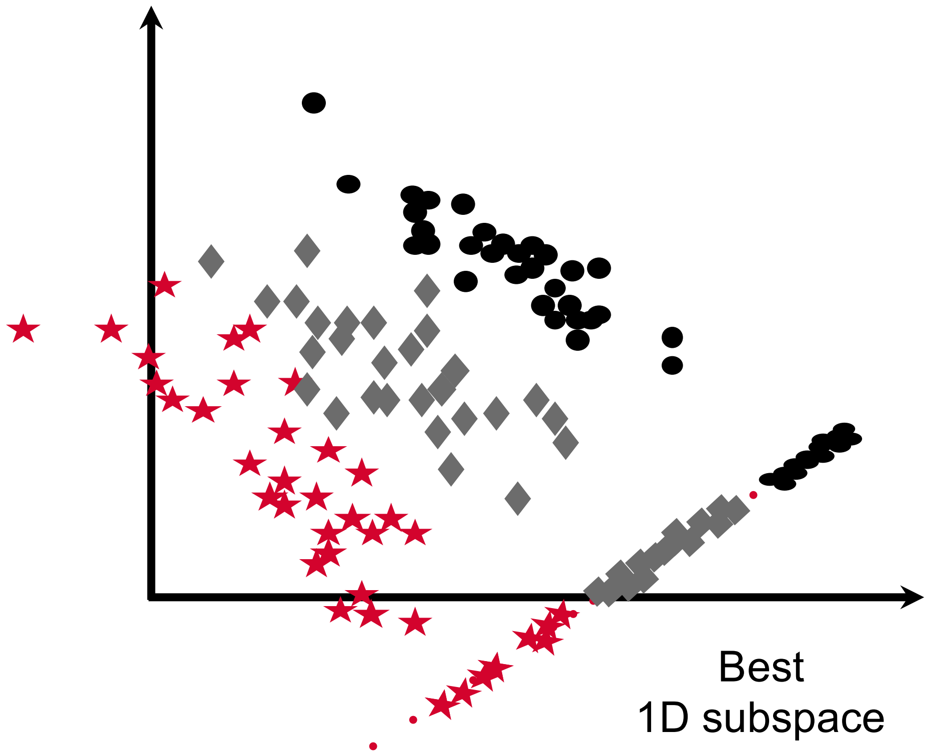
Understanding LDA



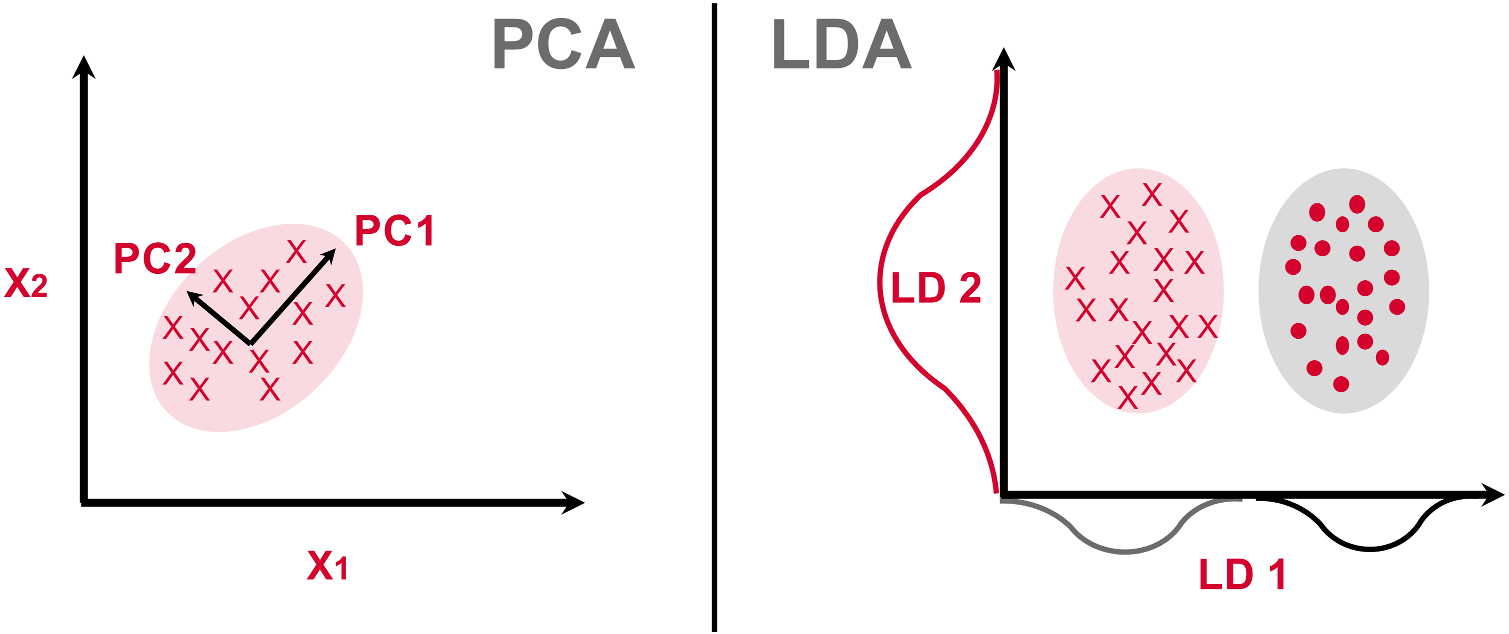
Choosing Axes



Choosing Axes



PCA vs. LDA

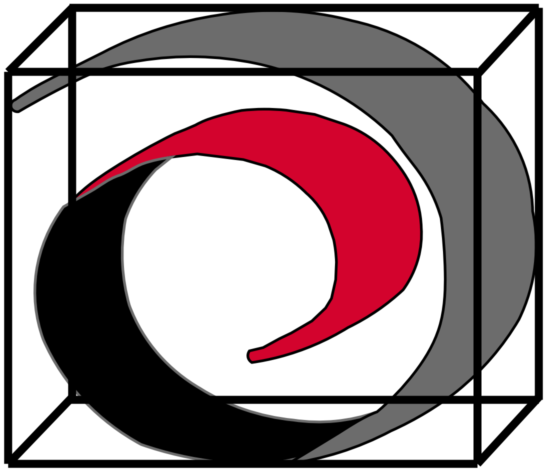


Choosing Manifold Learning

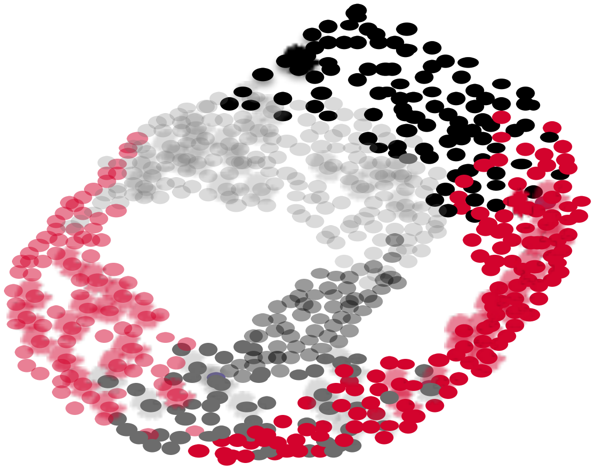
- Y not linearly related to X
- Very high dimensionality of X (e.g. pixel counts in image data)
- Sparse features, points are not dense clustered together in space
- Three-dimensional plots of indicate manifold shape

Manifold shape - a **simpler shape** in lower dimensions has been folded up to form a more complex shape in higher dimensions

Manifold Data

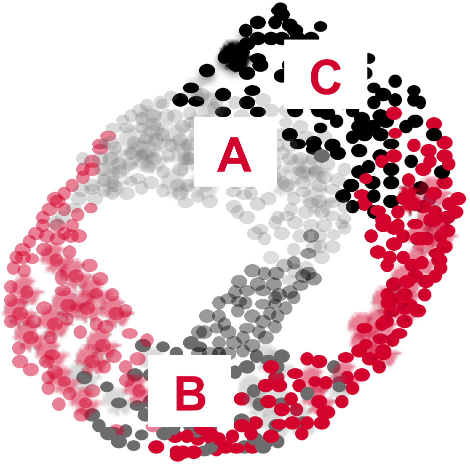


Manifold Hypothesis

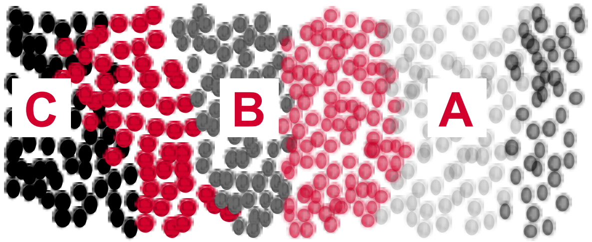


Many high-dimensional datasets can be easily unrolled
so that they lie along a much lower dimensional manifold

Manifold Learning



High-dimensional data



Unrolled to a simpler shape in low dimensions

Manifold Learning Techniques

MDS

Isomap

Kernel PCA

LLE

t-SNE

Multidimensional Scaling (MDS)

Aims to preserve pair-wise Euclidean distances between all points while reducing dimensionality. Some intuitive similarities to MSE regression in underlying math.

Isomap

Aims to preserve pair-wise Euclidean distances between neighboring points only (not all points) while reducing dimensionality; works out equivalent to preserving geodesic distance between all points.

Locally Linear Embedding

Expresses each point as centroid (weighted average) of nearest neighbors; then tries to maintain same weights upon conversion to new dimensions.

t-distributed Stochastic Neighbor Embedding (t-SNE)

Aims to keep similar points together and dissimilar points apart. First fits a Student-t probability distribution to the data, hence the name. Widely used in visualizing clusters.

Kernel PCA

First apply the kernel trick to map data into very high dimensional space. Then perform PCA to come down to lower-dimensional space.

Hands-on demos on dimensionality reduction

Polls

Poll 5

Each of the following techniques is used in manifold unrolling, with the exception of:

- Factor Analysis
- Locally Linear Embedding
- Multidimensional Scaling

Poll 5

Each of the following techniques is used in manifold unrolling, with the exception of:

- **Factor Analysis**
- Locally Linear Embedding
- Multidimensional Scaling

Poll 6

Isomap is a manifold learning technique that:

- Seeks to preserve geodesic distance between all points in the lower-dimensionality space
- Seeks to preserve pair-wise Euclidean distances between all points in the lower-dimensionality space

Poll 7

The primary difference between PCA and LDA is:

- PCA is a linear technique while LDA is a manifold learning technique
- PCA finds axes that maximize variance, LDA finds axes that maximize inter-class separation
- PCA is a manifold learning technique while LDA is a linear technique

Poll 7

The primary difference between PCA and LDA is:

- PCA is a linear technique while LDA is a manifold learning technique
- **PCA finds axes that maximize variance, LDA finds axes that maximize inter-class separation**
- PCA is a manifold learning technique while LDA is a linear technique

Poll 6

Isomap is a manifold learning technique that:

- **Seeks to preserve geodesic distance between all points in the lower-dimensionality space**
- Seeks to preserve pair-wise Euclidean distances between all points in the lower-dimensionality space