Introduction to statistical learning: supervised and unsupervised learning

prof. Janez Povh, Ph.D.

University in Ljubljana, Faculty of mechanical engineering

21st February 2017

Contents

What is SL?

1. Statistical learning (SL) is about (**deeper**) understanding of the data!
2. SL tries to answer:

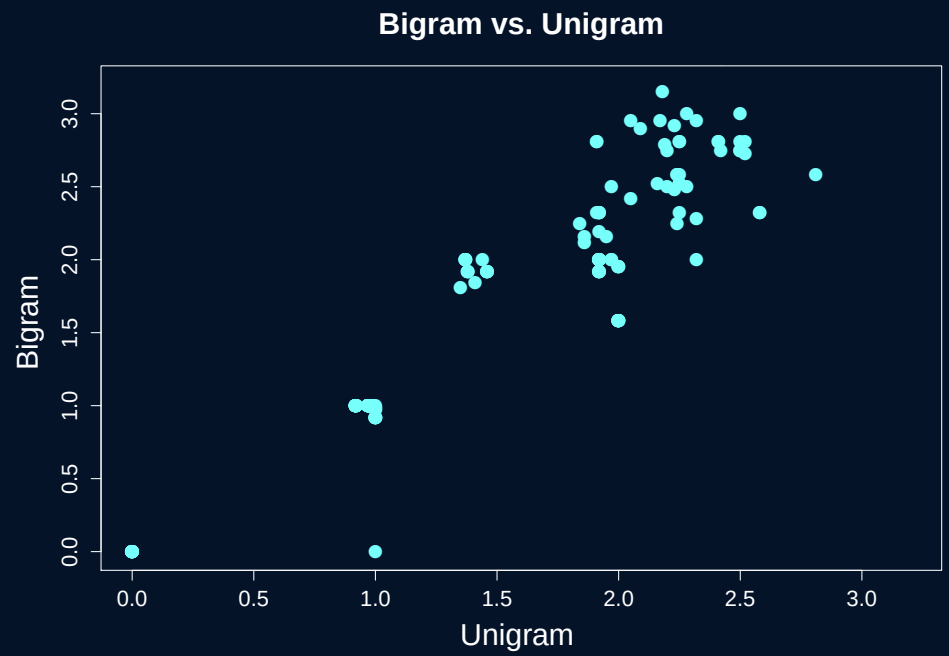
Question: SL

What are the hidden relations between the data instances and/or the data variables?

Example: regression

Example: SL

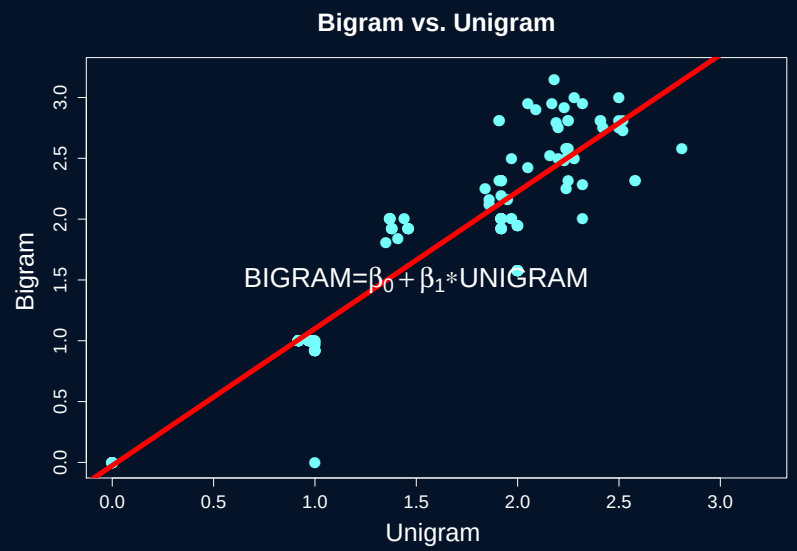
- dataset of songs;
- song's complexity: **UNIGRAM** and **BIGRAM** entropy.



Example: regression

Question: SL

What is the **relation** between BIGRAM and UNIGRAM?

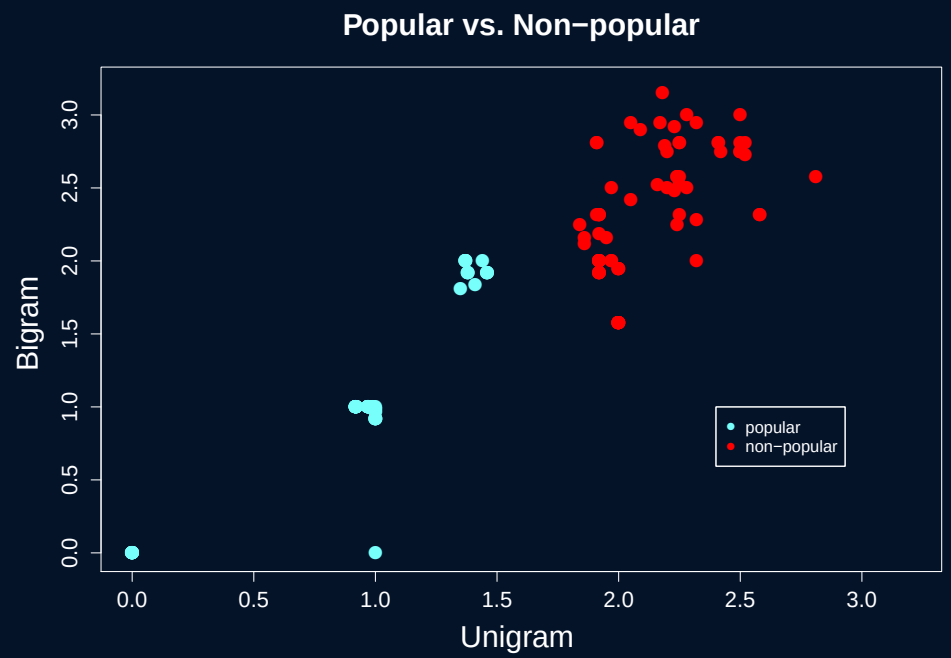


Example: classification

Example: SL

We have a dataset of songs. For each song we know:

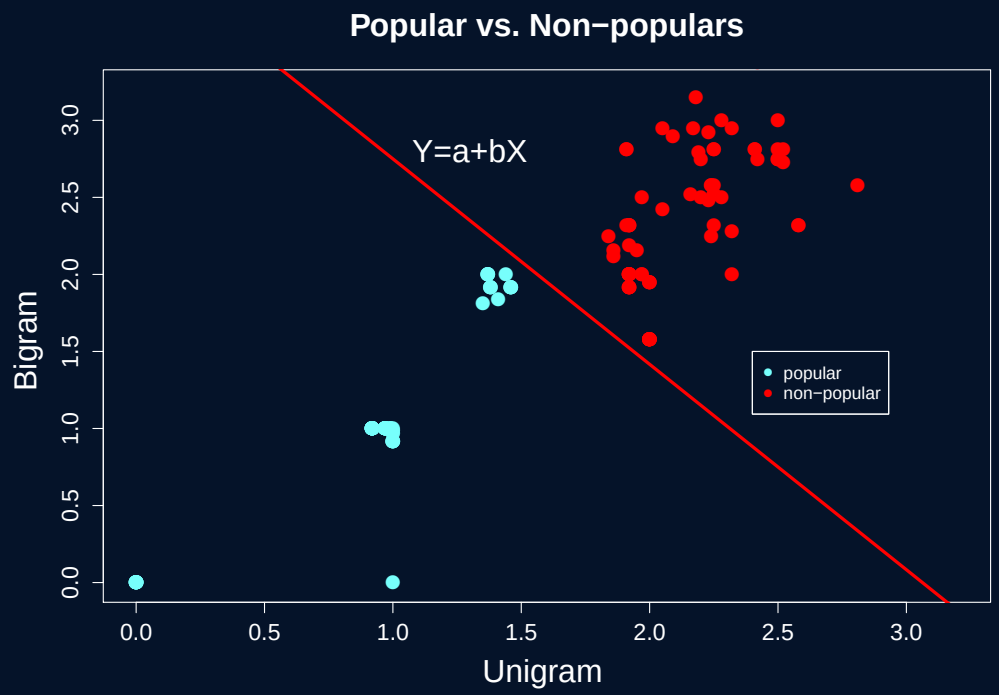
- song's complexity: **UNIGRAM** and **BIGRAM**.
- song's popularity: **popular** and **non-popular**.



Example: classification

Question: SL

Can we **predict** popularity based on BI-GRAM and UNI-GRAM entropy?



Supervised learning - **definition**

Definition: Supervised learning

- We **consider** variables $\mathbf{X}_1,\ \mathbf{X}_2, \cdots, \mathbf{X}_p$ (predictors) and $\mathbf{Y}$ (response);
- We **measured** $\mathbf{X}_1,\ \mathbf{X}_2, \cdots, \mathbf{X}_p$ and $\mathbf{Y}$ on given **(training)** set;
- We **look for** function $\mathbf{f}$ such that

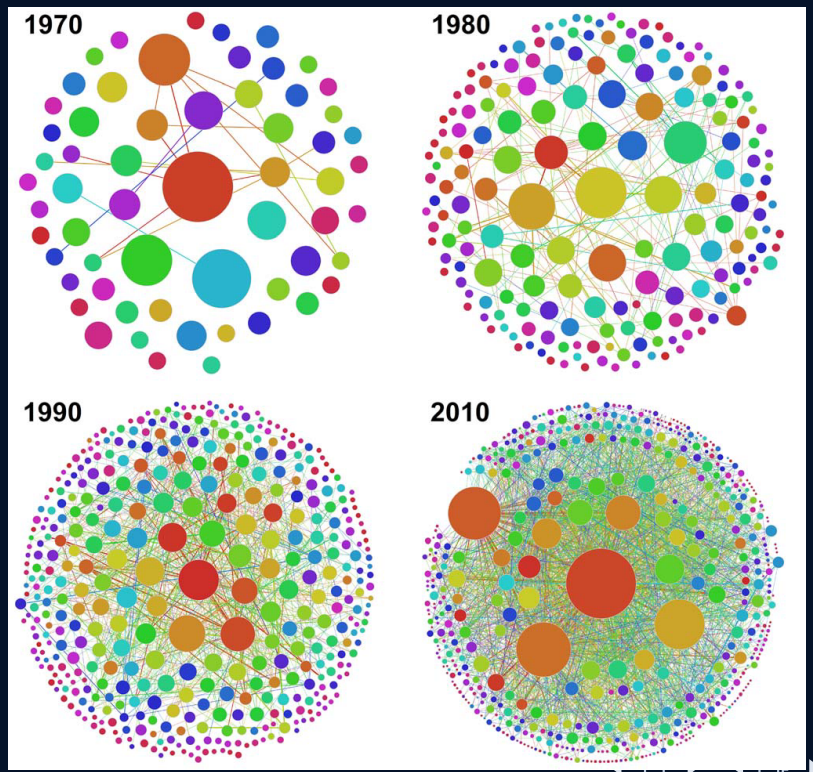
$$\mathbf{Y} = \mathbf{f}(\mathbf{X}_1,\ \mathbf{X}_2, \cdots, \mathbf{X}_p).$$

- **Examples:** classification, linear regression.

Example: clustering

Example

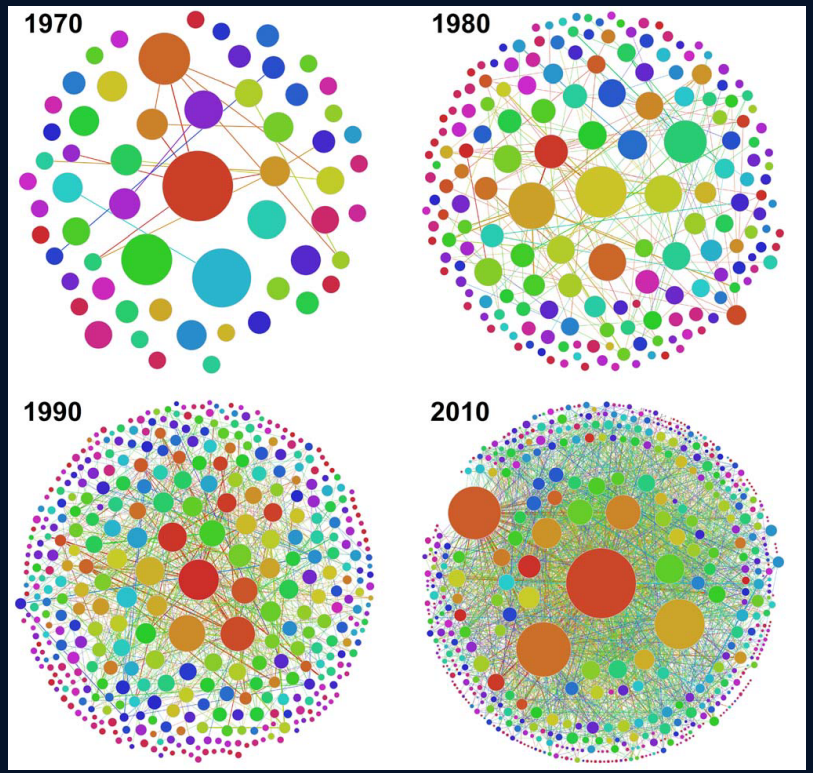
- Dataset of Slovenian scientists that published at least 1 paper in 1970-2015;
- Their collaboration in 1970, 1980, 1990 and 2010:



Example: clustering

Question

Can we **explain** the groups (communities) from the figure?

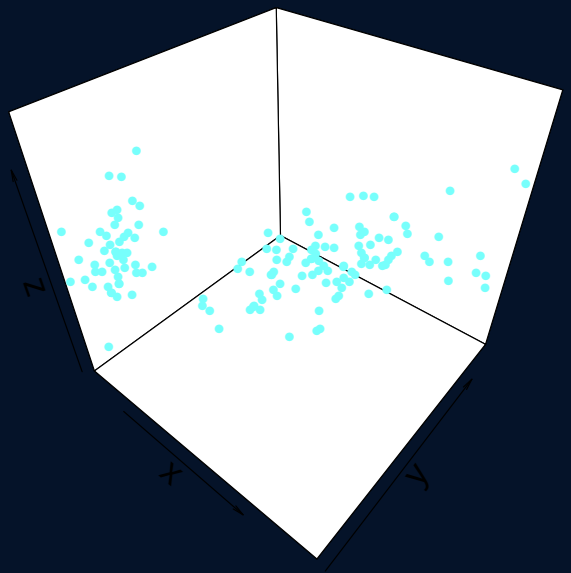


Example: dimension reduction

Example

- Dataset of cancer patients;
- p variables $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$ (**features**): sickness history, current status etc.
- How to **best** visualize these data in 2-dimensions?

3D scatter plots



Example: dimension reduction

Example

- Dataset of cancer patients;
- p variables $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$ (**features**): sickness history, current status etc.
- How to **best** visualize these data in 2-dimensions?

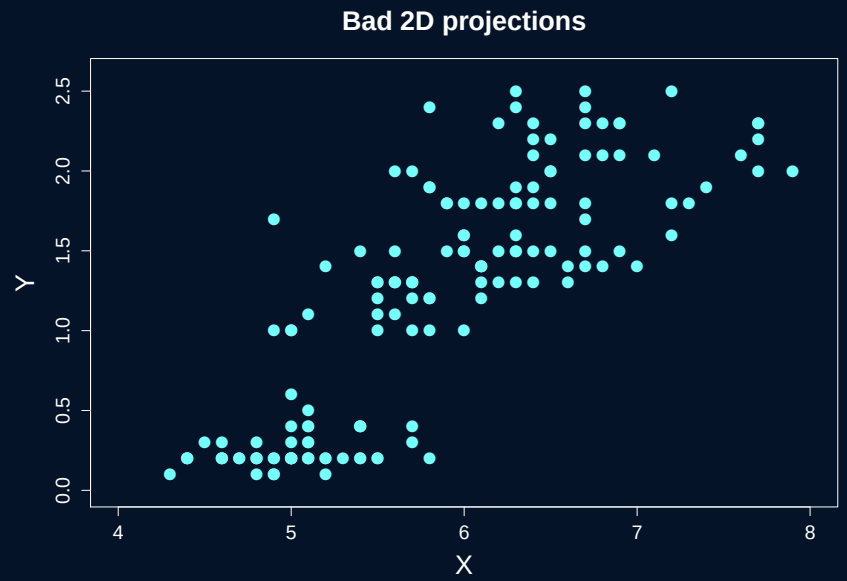


Figure: Scatter plot for 2D data

Example: dimension reduction

Example

- Dataset of cancer patients;
- p variables $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$ (**features**): sickness history, current status etc.
- How to **best** visualize these data in 2-dimensions?

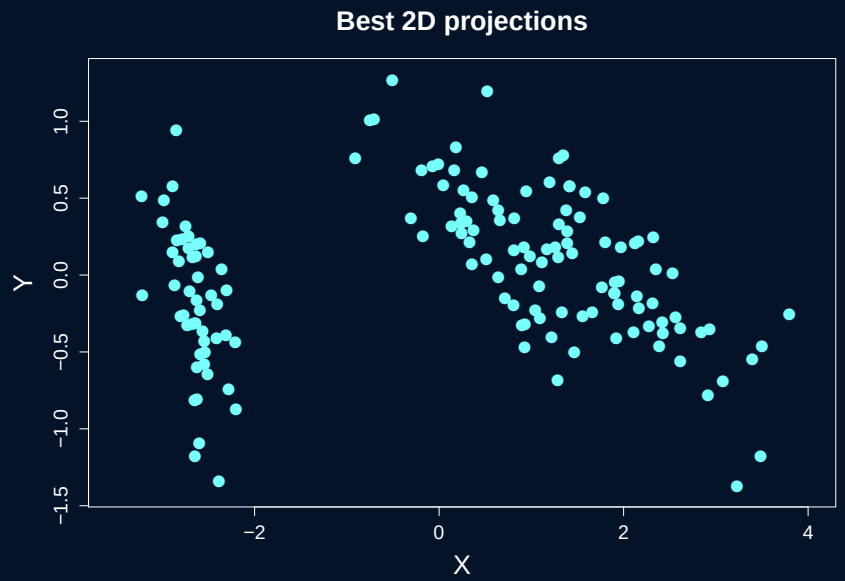


Figure: Scatter plot for 2D data

Unsupervised learning - **definition**

Definition: Unsupervised learning

- We **consider** variables $\mathbf{X}_1,\ \mathbf{X}_2, \cdots, \mathbf{X}_p$;
- We **measured** $\mathbf{X}_1,\ \mathbf{X}_2, \cdots, \mathbf{X}_p$ on given set;
- We **look for**
  - hidden groups of data instances (clusters) ($\Longrightarrow$ **CLUSTERING**) OR
  - **few** new variables enabling better visualization: ($\Longrightarrow$ **PRINCIPAL COMPONENT ANALYSIS**) OR
  - **few** new variables enabling better interpretation: ($\Longrightarrow$ **FACTOR ANALYSIS**).

Supervised vs. unsupervised learning - summary

	SUPERVISED LEARNING	UNSUPERVISED LEARNING
AVAILABLE DATA	$X_1, X_1, \ldots, X_p, Y$	$X_1, X_1, \ldots, X_p$
GROUND TRUTH	• We know the answer on available data • Results can be evaluated • Results can be compared	• Several solutions possible • No universal measure to compare solutions • Subjectivity