Unsupervised learning introduction

Brooks Paige

Weeks 7 and 8

Supervised learning

Most of the examples we've seen so far had datasets that looked like this:

$$\text{Inputs} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$$

$$\text{Labels} = \{y_1, \ldots, y_N\}$$

Supervised learning

Most of the examples we've seen so far had datasets that looked like this:

$$\text{Inputs} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$$
$$\text{Labels} = \{y_1, \ldots, y_N\}$$

- In **supervised** learning tasks, we have been trying to predict the labels y_i from the inputs $\mathbf{x}_i$

Supervised learning

Most of the examples we've seen so far had datasets that looked like this:

$$\text{Inputs} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$$
$$\text{Labels} = \{y_1, \dots, y_N\}$$

- In **supervised** learning tasks, we have been trying to predict the labels y_i from the inputs $\mathbf{x}_i$
- From a probabilistic perspective, this corresponds to trying to approximate the conditional distribution $p(y|\mathbf{x})$

Supervised learning

Most of the examples we've seen so far had datasets that looked like this:

$$\text{Inputs} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$$
$$\text{Labels} = \{y_1, \ldots, y_N\}$$

- In **supervised** learning tasks, we have been trying to predict the labels y_i from the inputs $\mathbf{x}_i$
- From a probabilistic perspective, this corresponds to trying to approximate the conditional distribution $p(y|\mathbf{x})$
- The inputs might have been complex and high dimensional, but the labels were mostly either real-valued, or discrete class labels

Unsupervised learning

For the next two weeks, we'll (mostly) be looking at datasets **with no labels**:

$$\text{Data} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}.$$

Unsupervised learning

For the next two weeks, we'll (mostly) be looking at datasets **with no labels**:

$$\text{Data} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}.$$

- In **unsupervised** learning tasks, we try to build a model for this data $\mathbf{x}_i \in \mathbb{R}^D$, which is possibly very high-dimensional.

Unsupervised learning

For the next two weeks, we'll (mostly) be looking at datasets **with no labels**:

$$\text{Data} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}.$$

- In **unsupervised** learning tasks, we try to build a model for this data $\mathbf{x}_i \in \mathbb{R}^D$, which is possibly very high-dimensional.
- From a probabilistic perspective, we assume the data is drawn from some **unknown** underlying distribution $p(\mathbf{x})$, which we will try to somehow characterize.

Density estimation

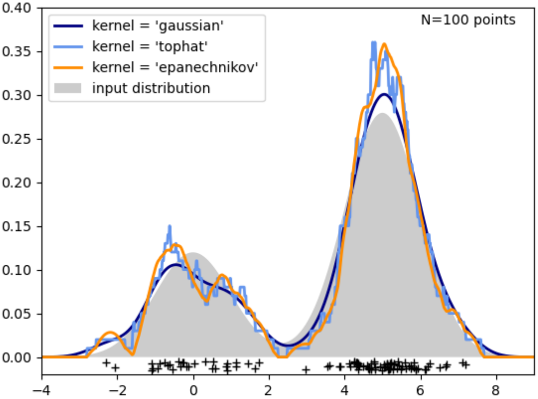
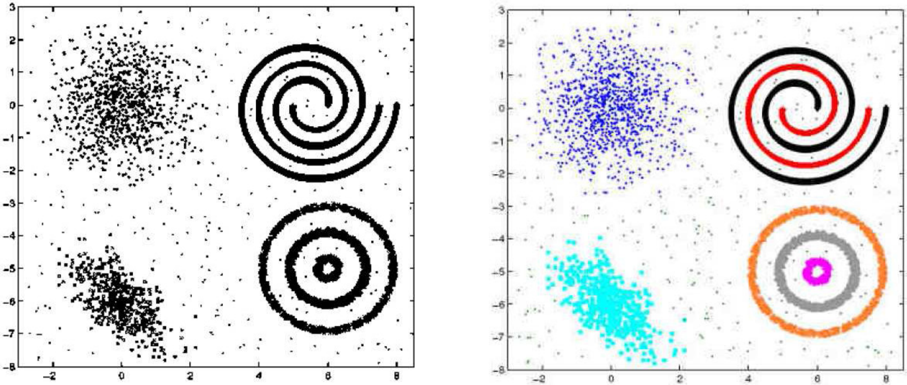


Figure: Scikit-learn KDE

Clustering



(a) Input data

(b) Desired clustering

Figure: Jain 2009

Clustering

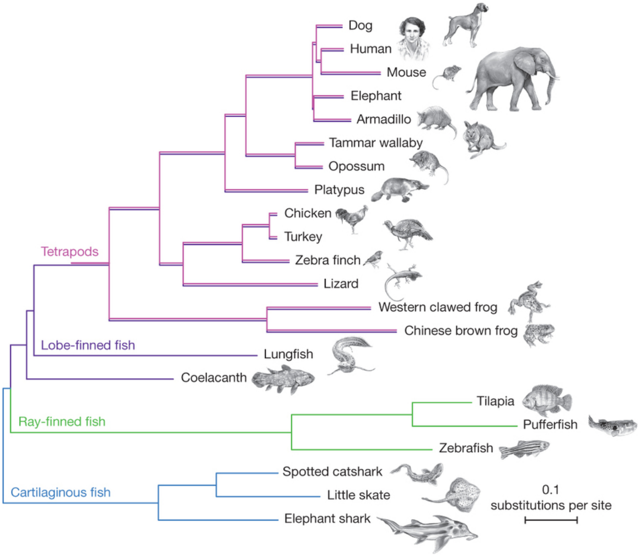


Figure: Amemiya et al., Nature (2013)

Mixture modelling

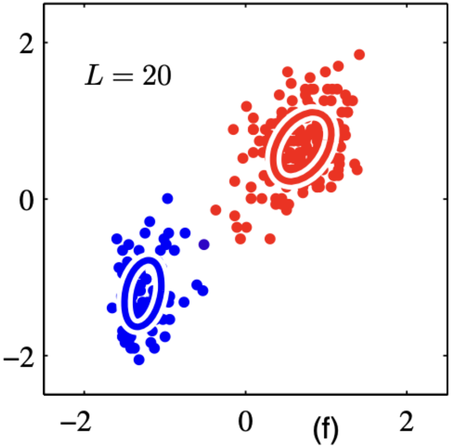


Figure: PRML

Dimensionality reduction

- **Clustering** explains the data by finding a **small number of modes** which account for much of the variation

Dimensionality reduction

- **Clustering** explains the data by finding a **small number of modes** which account for much of the variation

- **Dimensionality reduction** explains the data by finding a **low-dimensional subspace** or manifold which accounts for much of the variation

Dimensionality reduction

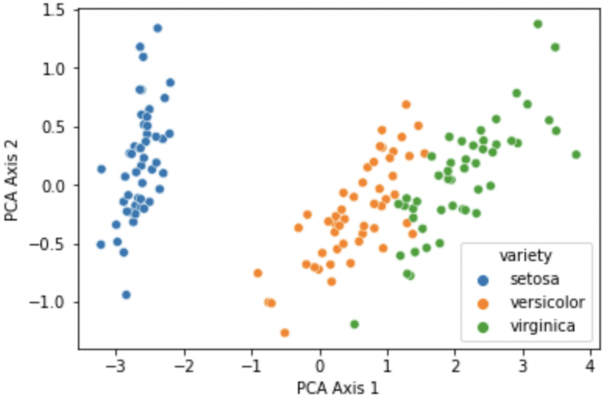
- **Clustering** explains the data by finding a **small number of modes** which account for much of the variation

- **Dimensionality reduction** explains the data by finding a **low-dimensional subspace** or manifold which accounts for much of the variation

Many reasons we may want to do this!

- visualization and human interpretation
- removing spurious features, including as pre-processing for other algorithms
- uncovering meaningful latent variables

Principal Components Analysis



Visualization

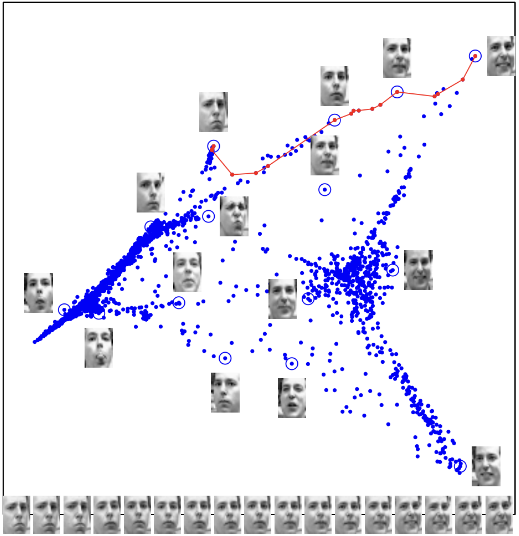


Figure: Roweis & Saul (2000)

Matrix factorization

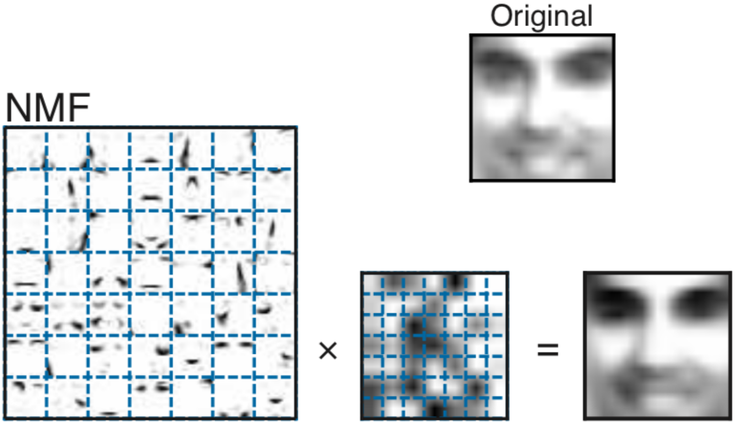
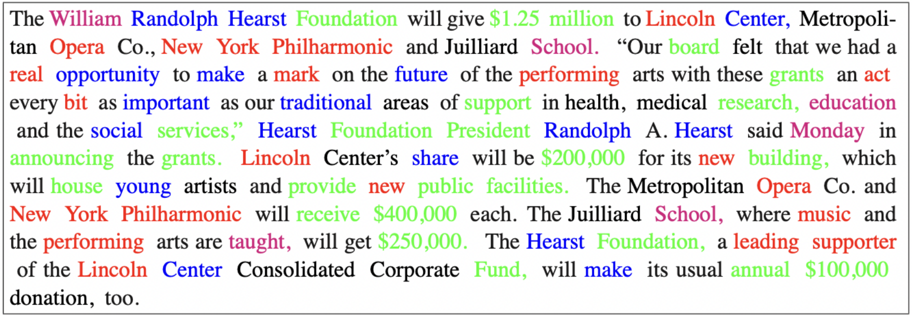


Figure: ESL, Ch. 14

Topic modelling

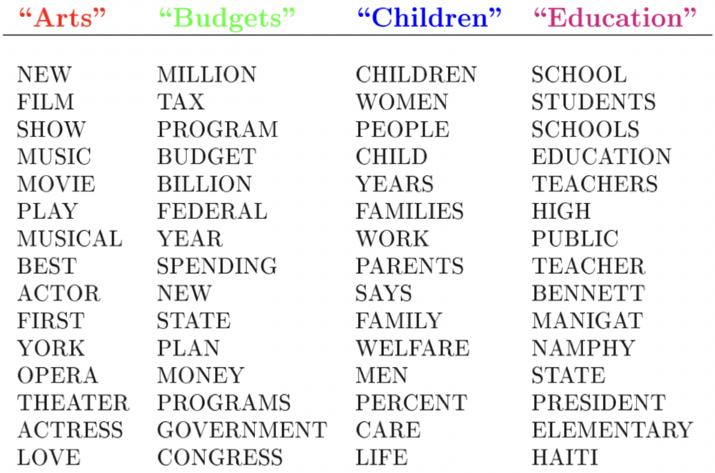


The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Figure: Blei et al. (2003)

Topic modelling



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Figure: Blei et al. (2003)

Evaluation?

How do we evaluate performance of these methods?

(Is looking at the loss enough?)

Evaluation?

How do we evaluate performance of these methods?

(Is looking at the loss enough?)

- **Probabilistic models**: look at the likelihood

Evaluation?

How do we evaluate performance of these methods?

(Is looking at the loss enough?)

- **Probabilistic models**: look at the likelihood
- **Clustering**: compare with known labels...? Check for overlap...?
  - Best option: look at performance when used on downstream tasks

Evaluation?

How do we evaluate performance of these methods?

(Is looking at the loss enough?)

- **Probabilistic models**: look at the likelihood
- **Clustering**: compare with known labels...? Check for overlap...?
  - Best option: look at performance when used on downstream tasks
- **Visualization**: user studies...?