Introduction to Machine Learning

Lecture 6: Conclusion

Alexis Zubiolo
alexis.zubiolo@gmail.com

Data Science Team Lead @ Adcash

December 8, 2016

Outline

This lecture includes:

- A **model selection** lab
- A few aspects of machine learning we have not mentioned
  - **Feature engineering**
  - **Dimensionality reduction**
- Feedback and final Q&A

Dimensionality reduction

Dimensionality reduction

When working on real-world data sets, the data is not always clean and ready to use.

Dimensionality reduction

When working on real-world data sets, the data is not always clean and ready to use.

For example, in some data sets, one may encounter the following issues:

- There are too many samples
- There are too many features
- Some of the features bring no information

Dimensionality reduction

When working on real-world data sets, the data is not always clean and ready to use.

For example, in some data sets, one may encounter the following issues:

- There are too many samples
- There are too many features
- Some of the features bring no information

Hence, there is a need to pre-process the data.

Dimensionality reduction

In case there are too many samples, a simple solution is to apply subsampling, *e.g.* just take into account 10% of the samples.

- Recall the mini-batch k-means
- Be careful with the **class balance**
  - Imbalanced classes: **subsample the majority class**
  - Little to no imbalance: **Stratified sampling**

Dimensionality reduction

In case there are too many samples, a simple solution is to apply subsampling, *e.g.* just take into account 10% of the samples.

- Recall the mini-batch k-means
- Be careful with the **class balance**
  - Imbalanced classes: **subsample the majority class**
  - Little to no imbalance: **Stratified sampling**

It is also important to use common sense and expert knowledge when possible:

- Some variables might be intuitively meaningless to solve the ML problem
- ML often aims at mimicking/automatic a human logic

Dimensionality reduction with PCA

PCA = Principal Component Analysis

Dimensionality reduction with PCA

PCA = Principal Component Analysis

Rough idea:

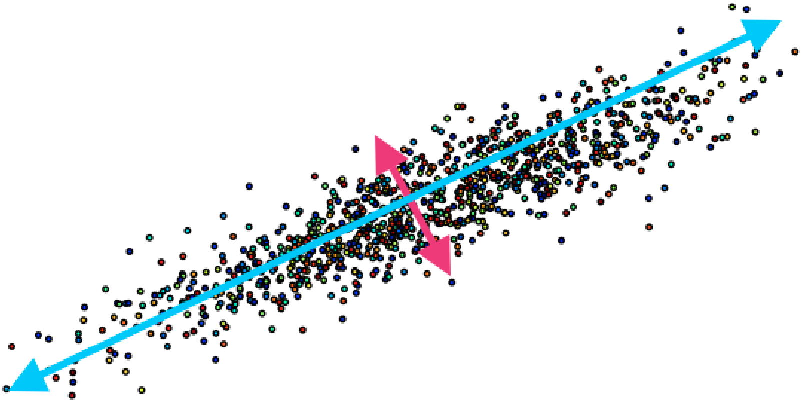
- Find high variance axes
- Select the k axes with the highest variances
- Project the data on these axes

Dimensionality reduction with PCA

PCA = Principal Component Analysis

Rough idea:

- Find high variance axes
- Select the k axes with the highest variances
- Project the data on these axes



Feature engineering

Feature engineering

Feature engineering is a key component of machine learning.

Feature engineering

Feature engineering is a key component of machine learning.

Sometimes, on the same data, most algorithm would give roughly the same accuracy (or any other metric). In this case, feature engineering could be the best solution to improve classification/regression results.

Feature engineering

Feature engineering is a key component of machine learning.

Sometimes, on the same data, most algorithm would give roughly the same accuracy (or any other metric). In this case, feature engineering could be the best solution to improve classification/regression results.

Feature engineering is often **data-dependent**: You won't use the same features from text data and from images or videos.

Feature engineering in text analysis

There are several challenges when dealing with text data:

- Mining text can lead to a huge amount of data to process
- Not all the data is relevant
- Texts can be of different size
- ...

Feature engineering in text analysis

There are several challenges when dealing with text data:

- Mining text can lead to a huge amount of data to process
- Not all the data is relevant
- Texts can be of different size
- ...

Hence, there is **a need to preprocess** it before giving it to any ML algorithm.

Feature engineering in text analysis with TF-IDF

Given terms $t \in T$ and documents $d \in D$, we can define:

$$\text{TF}(t, d) = \text{Frequency of term } t \text{ in the document } d$$

$$\text{IDF}(t, D) = \log \left(\frac{|D|}{|\{d \in D, t \in d\}|} \right)$$

(there are alternative formulas)

Feature engineering in text analysis with TF-IDF

Given terms $t \in T$ and documents $d \in D$, we can define:

$$\text{TF}(t, d) = \text{Frequency of term } t \text{ in the document } d$$

$$\text{IDF}(t, D) = \log\left(\frac{|D|}{|\{d \in D, t \in d\}|}\right)$$

(there are alternative formulas)

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

Feature engineering in text analysis with TF-IDF

Given terms $t \in T$ and documents $d \in D$, we can define:

$$\text{TF}(t, d) = \text{Frequency of term } t \text{ in the document } d$$

$$\text{IDF}(t, D) = \log\left(\frac{|D|}{|\{d \in D, t \in d\}|}\right)$$

(there are alternative formulas)

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

Exercice: Cf. Board

Feature engineering in text analysis with TF-IDF

Given terms $t \in T$ and documents $d \in D$, we can define:

$$\text{TF}(t, d) = \text{Frequency of term } t \text{ in the document } d$$

$$\text{IDF}(t, D) = \log\left(\frac{|D|}{|\{d \in D, t \in d\}|}\right)$$

(there are alternative formulas)

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

Exercice: Cf. Board Note: We can go further, *e.g.* with *n*-grams.

Feature analysis in image processing

In images, we often want to detect interest points such as **edges** and **corners**.

Feature analysis in image processing

In images, we often want to detect interest points such as **edges** and **corners**.

SIFT = Scale-Invariant Feature Transform. Invariant to:

- Rotation
- Difference scales
- Affine transformation
- (Affine) intensity change

Feature analysis in image processing

In images, we often want to detect interest points such as **edges** and **corners**.

SIFT = Scale-Invariant Feature Transform. Invariant to:

- Rotation
- Difference scales
- Affine transformation
- (Affine) intensity change

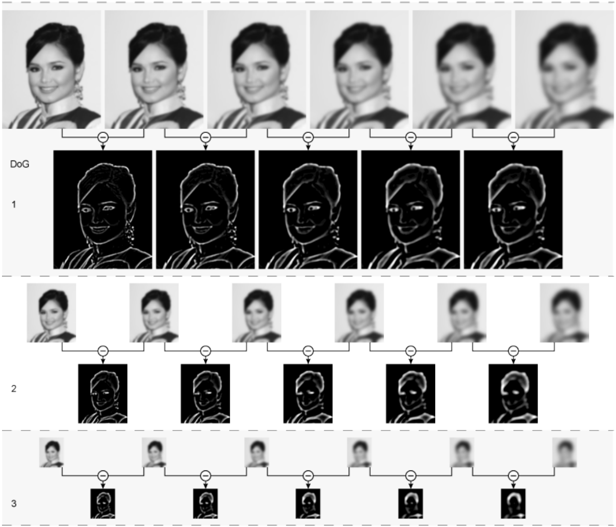
SIFT is patented and cannot be used in all situations: There exist alternatives based on the same idea such as SURF (Speeded-Up Robust Features)

SIFT algorithm

Key steps:

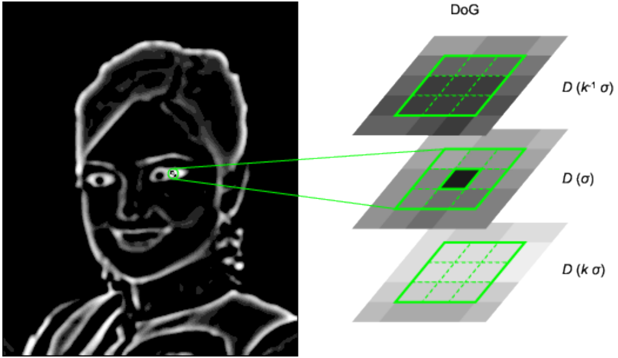
- Detect extrema at different scale by difference of gaussians
- Detect interest points in the image
- Assign orientations to create SIFT features

SIFT: Difference of Gaussians (DoG)



SIFT: Extrema detection

Interest points are among the local extrema in the $3 \times 3 \times 3$ neighborhood:



SIFT: Extrema detection

Post-processing:

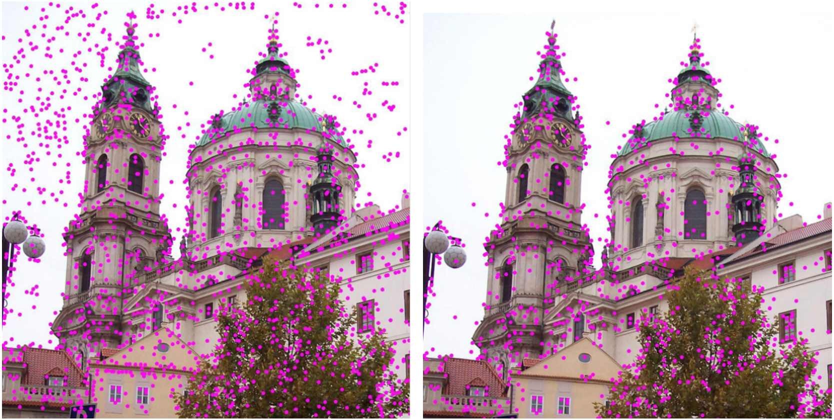
- Remove low-contrast points
- Remove points on the edges

SIFT: Extrema detection

Post-processing:

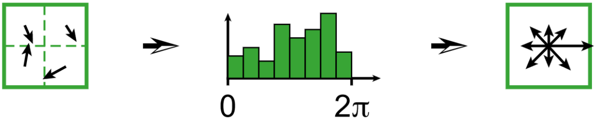
- Remove low-contrast points
- Remove points on the edges

Before/after:

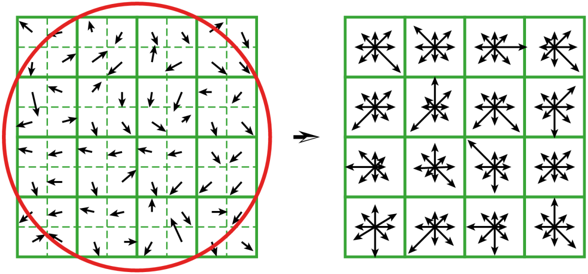


SIFT: Orientation assignment

For each interest point, compute the orientation histogram:



Do it for a neighborhood of the interest point:



SIFT applications

SIFT has other applications, such as aligning images, creating panoramas, video tracking, ...

SIFT applications

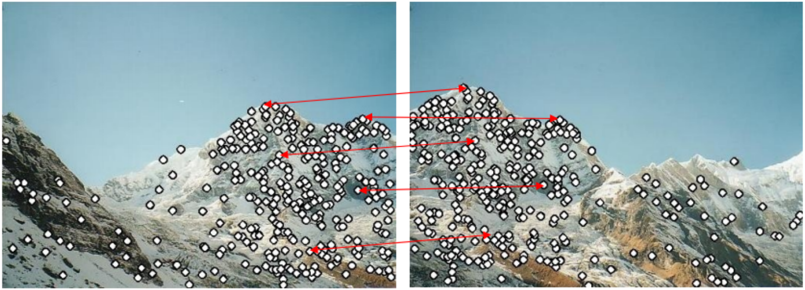
SIFT has other applications, such as aligning images, creating panoramas, video tracking, . . .

YouTube video example

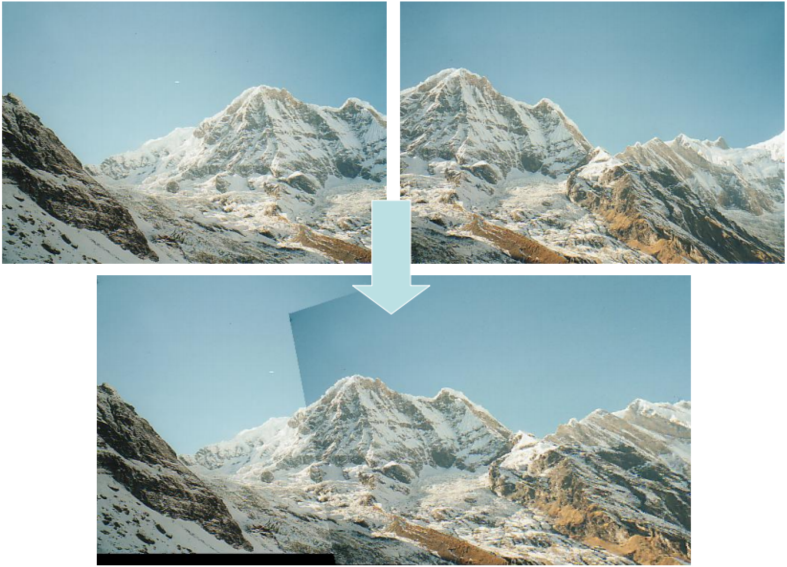
SIFT applications

SIFT has other applications, such as aligning images, creating panoramas, video tracking, ...

YouTube video example



SIFT illustration: Panorama



Conclusion of the conclusion

ML has plenty of applications as we saw

Conclusion of the conclusion

ML has plenty of applications as we saw

There's **no general-purpose solution** (at least for now): It is important to **look at the data** and **adapt to the situation**.

Conclusion of the conclusion

ML has plenty of applications as we saw

There's **no general-purpose solution** (at least for now): It is important to **look at the data** and **adapt to the situation**.

There are plenty of libraries and examples available online... Don't hesitate to play with it!

Thank you! Questions?
Any feedback?