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 mentionned
  - **Feature engineering**
  - **Dimensionality reduction**
- Feedback and final Q&A

Dimensionality reduction

Dimensionality reduction

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

Dimensionality reduction

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

For example, in some datasets, 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 datasets, the data is not always clean and ready to use.

For example, in some datasets, 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 with PCA

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: TF-IDF

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: TF-IDF

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 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. Invariance 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. Invariance 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 applications

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

SIFT applications

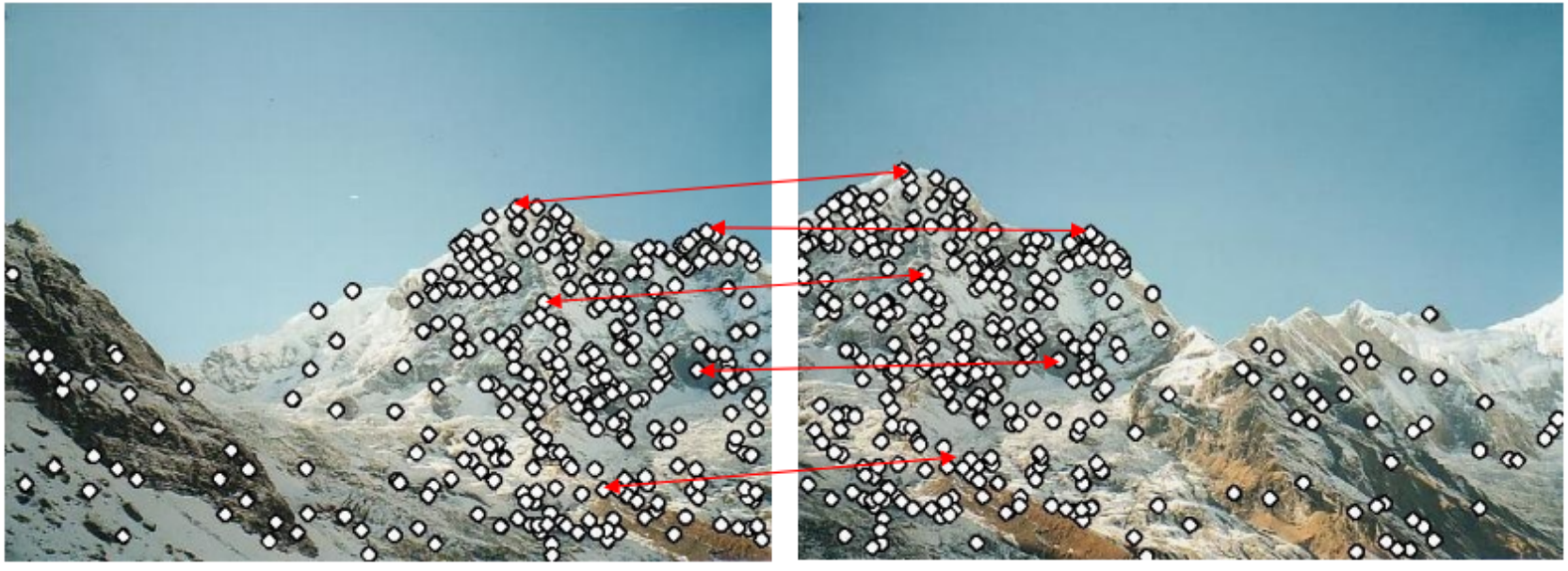
SIFT has other applications, such as aligning images, creating panoramas, video tracking, . . .

YouTube video example

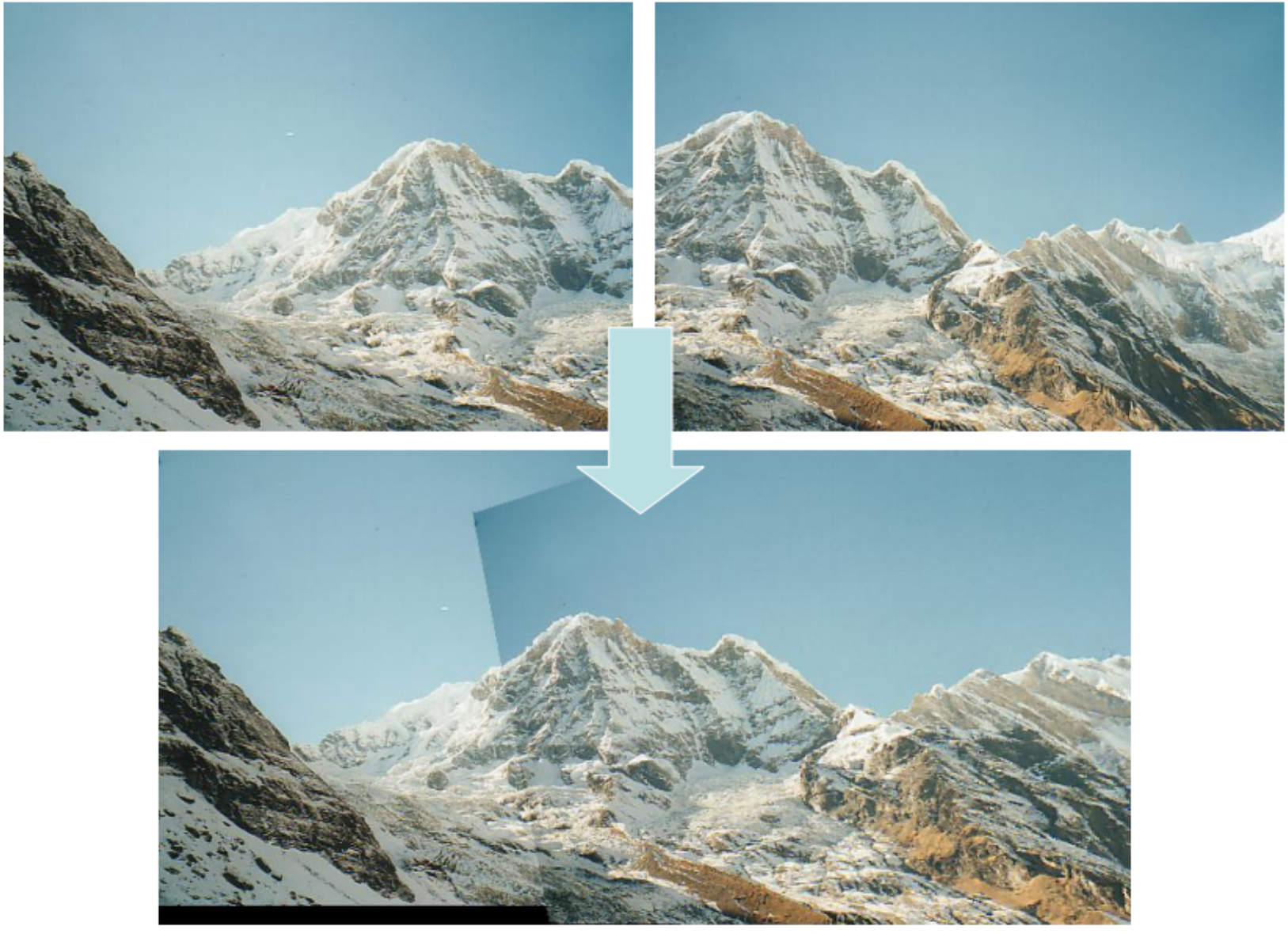
SIFT applications

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

YouTube video example



SIFT illustration



Conclusion of the conclusion

Thank you! Questions?
Any feedback?