Mississauga Big Data Meetup 26

-Kumar V

May 03, 2017

www.meetup.com/Mississauga-Big-Data-Analytics-Meetup/

Agenda:

Choose a Dataset to work on, for the next few weeks.

6 lines of expertise in Big Data Analytics

- Understanding of Business Processes that Produce Big Data. Understanding of problems/questions one could ask on such data (Business Domain Expertise)
- Understanding Algorithms, the math/stat backing them up, that could be applied on such data (Machine Learning Expertise).
- Big Data Architecting: Understanding of architectures that could be used to solve the problem (Big Data Architecture/Technology Stack Expertise)

- Understanding of the physical infrastructure and the cluster administration, security protocols, etc. (Big Data Infrastructure Administration)
- Understanding of Data Governance: setting up Data Stewards, processes for carrying out Data Access Administration, etc. (Data Governance)
- Application Development skills: Many different languages, many API, functional programming paradigms, etc. (Application Development)

My suggestions:

https://www.kaggle.com/dalpozz/creditcardfraud

https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city

https://www.kaggle.com/hhs/health-insurance-marketplace

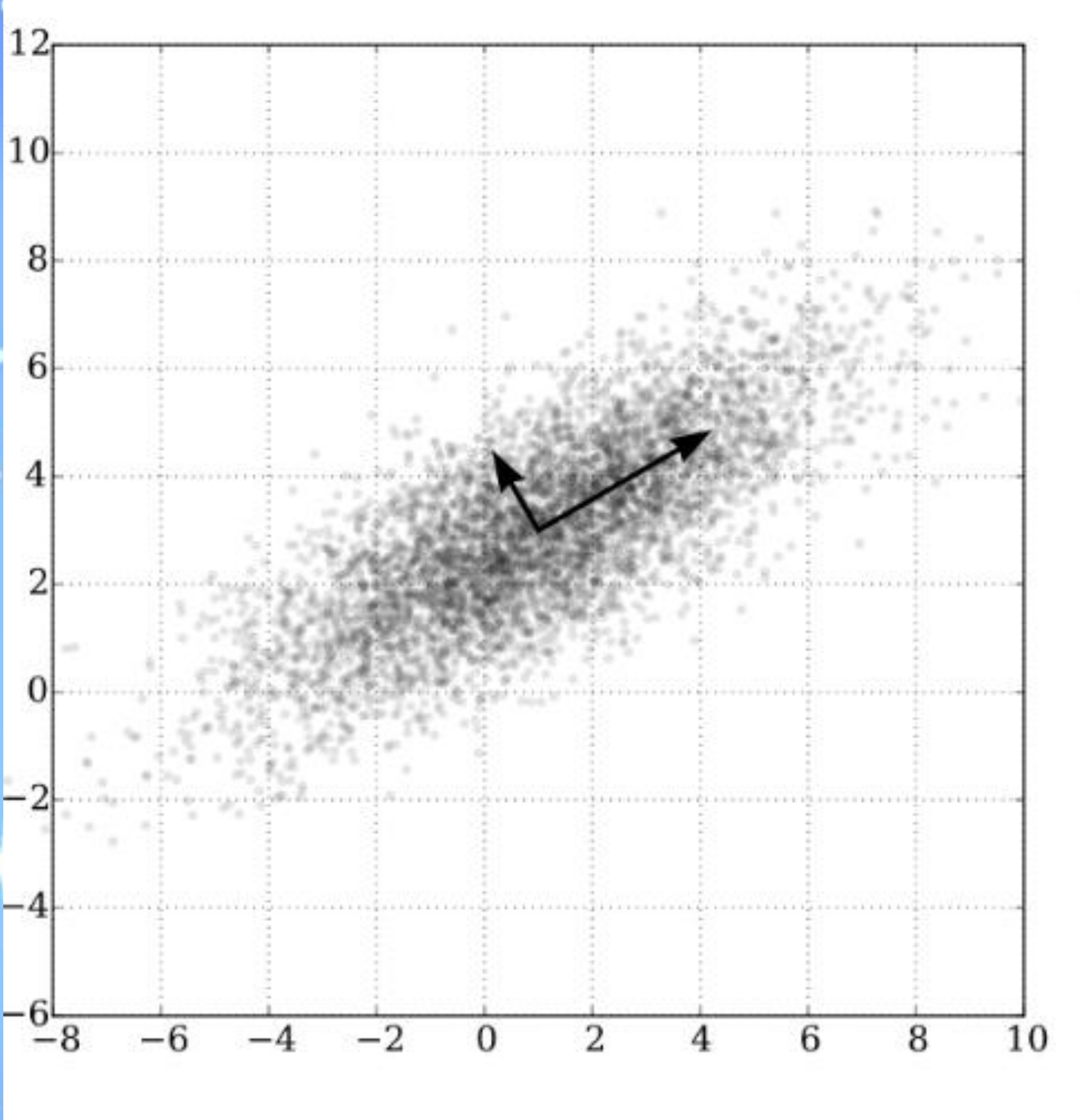
PCA

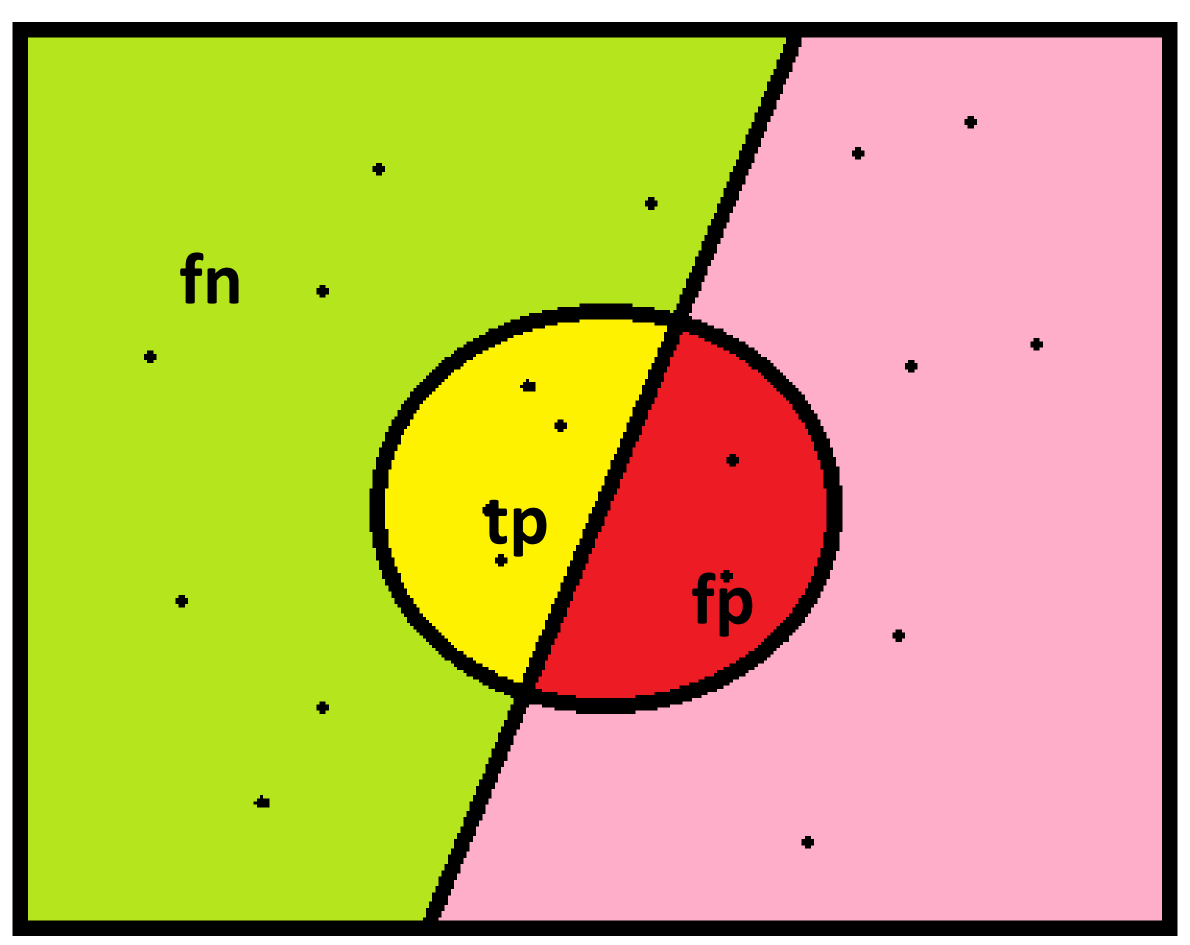
Wikipedia:

Principal component analysis (**PCA**) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (or sometimes, principal modes of variation).

Wikipedia:

- PCA can be thought of as fitting an *n*-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component. If some axis of the ellipsoid is small, then the variance along that axis is also small, and by omitting that axis and its corresponding principal component from our representation of the dataset, we lose only a commensurately small amount of information.


12
10
8
6
4
2
0
−2
−4
−6
−8
−6
−4
−2
0
2
4
6
8
10



tp = true positives
fp = false positives
fn = false negatives

$$\text{Precision} = \frac{tp}{tp + fp} \qquad \text{Recall} = \frac{tp}{tp + fn}$$

- In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

	Credit Card Fraud Dataset	Uber Pickups Dataset	Health Insurance Marketplace
Business Domain Expertise	---	---	---
Machine Learning Expertise	Yes	Yes	Yes
Big Data Arch / Technology Stack Expertise	May be	May be	May be
Big Data Deployment Administration	May be	May be	May be
Data Governance	---	---	---
Application Development Skills	Yes	Yes	Yes

Any

Questions

Or

Comments?