Tutorial Machine Learning in Python

Derek Harter

universitätbonn

GK Bionik Tutorial 2012

Outline

Introduction to Python

Unsupervised Learning

- PCA
- k-Means

Supervised Learning

- Linear Regression
- Classification
- Logistic Regression
- k Nearest Neighbors

Outline

universitätbonn

Introduction to Python

Unsupervised Learning

- PCA
- k-Means

Supervised Learning

- Linear Regression
- Classification
- Logistic Regression
- k Nearest Neighbors

A Short Introduction to Python

universitätbonn

- Please log in, using:

  Username gkbionik

  Password tut0rial (with a zero instead of the "o"!)

Outline

Introduction to Python

Unsupervised Learning

- PCA
- k-Means

Supervised Learning

- Linear Regression
- Classification
- Logistic Regression
- k Nearest Neighbors

Outline

Introduction to Python

Unsupervised Learning

- PCA
- k-Means

Supervised Learning

Motivation: Exploring High-Dim Data

- Let's say you do an experiment.
- You vary very few variables, and measure many different outcome variables.
- In our example, we change one variable, but measure four.

varied:

```
Y=[0,1,2,1,1,0,2,0,...]
```

observed:

```
X =
[[ 5.1   3.5   1.4   0.2]
 [ 4.9   3.    1.4   0.2]
 [ 4.7   3.2   1.3   0.2]
 [ 4.6   3.1   1.5   0.2]
 [ 5.    3.6   1.4   0.2]
 [ 5.4   3.9   1.7   0.4]
 [ 4.6   3.4   1.4   0.3]
 [ 5.    3.4   1.5   0.2]
 [ 4.4   2.9   1.4   0.2]
 ...
 [ 4.8   3.4   1.6   0.2]
 [ 4.8   3.    1.4   0.1]
 [ 4.3   3.    1.1   0.1]]
```

Motivation: Exploring High-Dim Data

- Let's say you do an experiment.
- You vary very few variables, and measure many different outcome variables.
- In our example, we change one variable, but measure four.
- You'd suspect there is a simple low-dimensional structure hidden in these four dimensions.

varied:

```
Y=[0,1,2,1,1,0,2,0,...]
```

observed:

```
X =
[[ 5.1  3.5  1.4  0.2]
 [ 4.9  3.   1.4  0.2]
 [ 4.7  3.2  1.3  0.2]
 [ 4.6  3.1  1.5  0.2]
 [ 5.   3.6  1.4  0.2]
 [ 5.4  3.9  1.7  0.4]
 [ 4.6  3.4  1.4  0.3]
 [ 5.   3.4  1.5  0.2]
 [ 4.4  2.9  1.4  0.2]
 ...
 [ 4.8  3.4  1.6  0.2]
 [ 4.8  3.   1.4  0.1]
 [ 4.3  3.   1.1  0.1]]
```

Plotting the Data

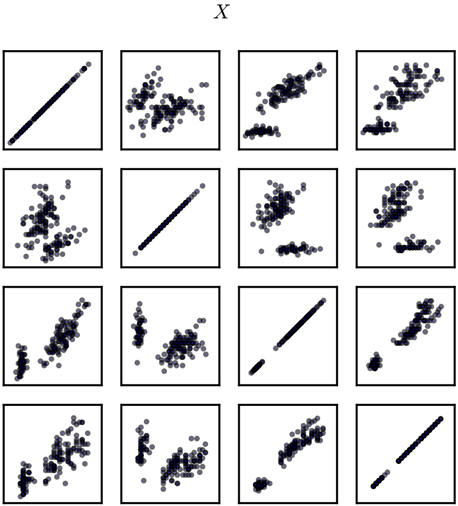
universitätbonn

- Looking at numbers is boring.

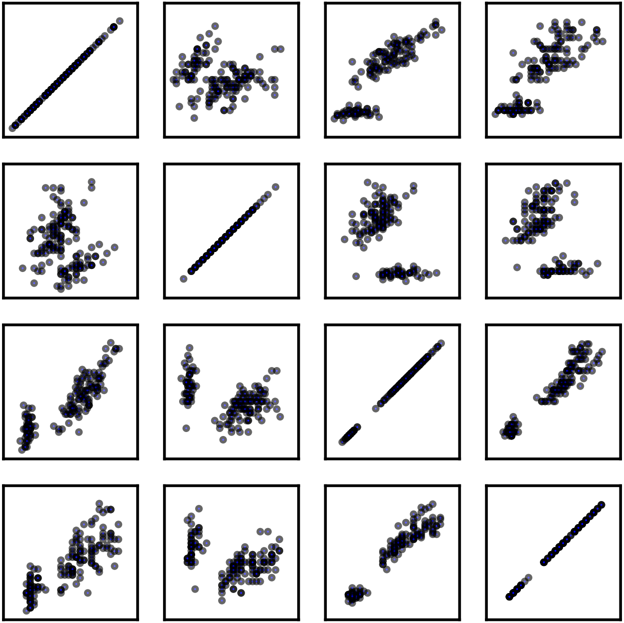
```
X =
[[ 5.1  3.5  1.4  0.2]
 [ 4.9  3.   1.4  0.2]
 [ 4.7  3.2  1.3  0.2]
 [ 4.6  3.1  1.5  0.2]
 [ 5.   3.6  1.4  0.2]
 [ 5.4  3.9  1.7  0.4]
 [ 4.6  3.4  1.4  0.3]
 [ 5.   3.4  1.5  0.2]
 [ 4.4  2.9  1.4  0.2]
 ...
 [ 4.8  3.4  1.6  0.2]
 [ 4.8  3.   1.4  0.1]
 [ 4.3  3.   1.1  0.1]]
```

Plotting the Data

- Looking at numbers is boring.
- 4 dimensions can be projected make 16 pairs



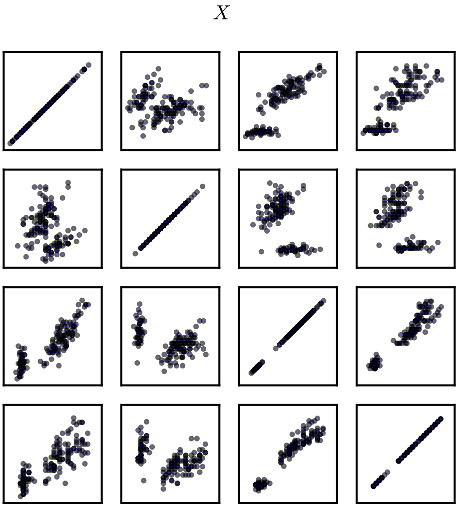
X



Plotting the Data

universitätbonn

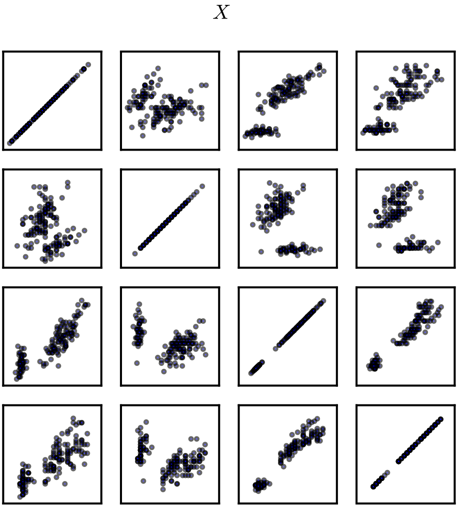
1. Which one of those projections is good?



Plotting the Data

universitätbonn

1. Which one of those projections is good?
2. Are there other, possibly better projections?

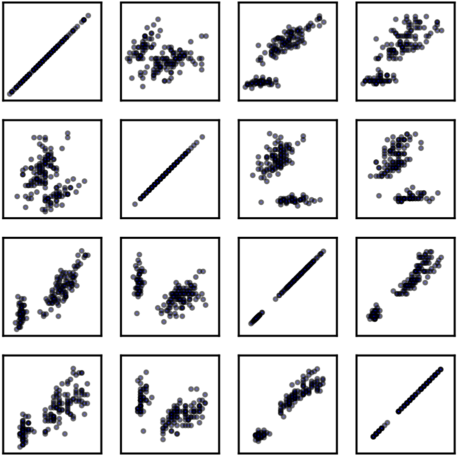


Plotting the Data

universitätbonn

1. Which one of those projections is good?
2. Are there other, possibly better projections?
3. Which variables are involved in the best projections?

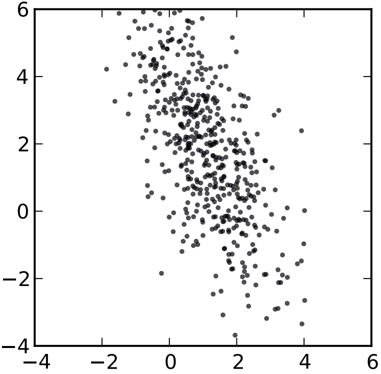
X



Principal Component Analysis

universitätbonn

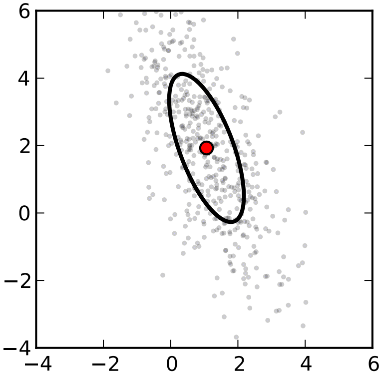
- In image on right, what is the "most important axis"?



Principal Component Analysis

universitätbonn

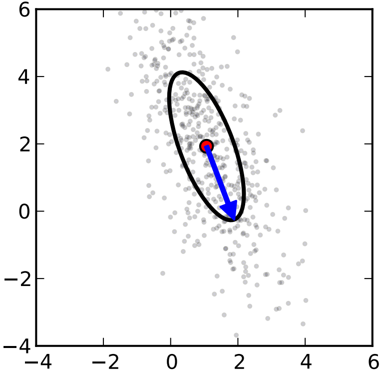
- In image on right, what is the "most important axis"?
- PCA models the data as a (multi-dimensional) ellipse



Principal Component Analysis

universitätbonn

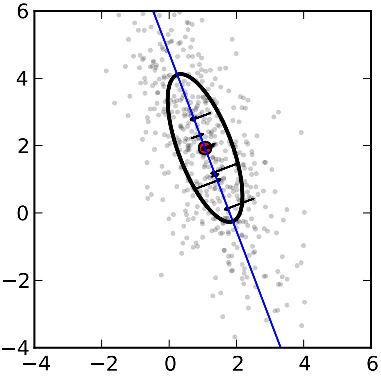
- In image on right, what is the "most important axis"?
- PCA models the data as a (multi-dimensional) ellipse
- PCA finds direction with largest variance (=diameter)



Principal Component Analysis

universitätbonn

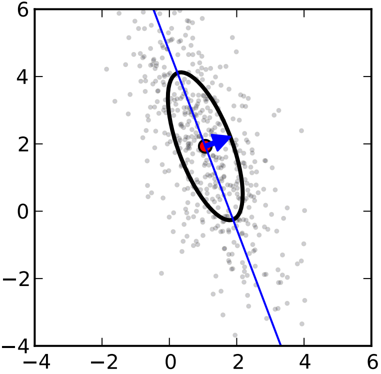
- In image on right, what is the "most important axis"?
- PCA models the data as a (multi-dimensional) ellipse
- PCA finds direction with largest variance (=diameter)
- First coordinate is projection onto this direction



Principal Component Analysis

universitätbonn

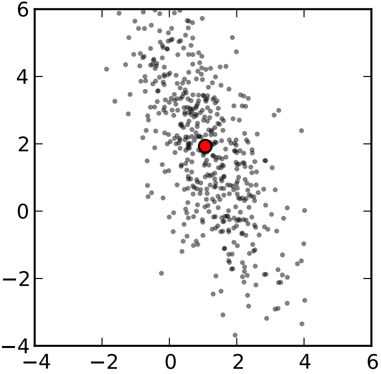
- In image on right, what is the "most important axis"?
- PCA models the data as a (multi-dimensional) ellipse
- PCA finds direction with largest variance (=diameter)
- First coordinate is projection onto this direction
- Continue with second, orthogonal axis...



Principal Component Analysis

universitätbonn

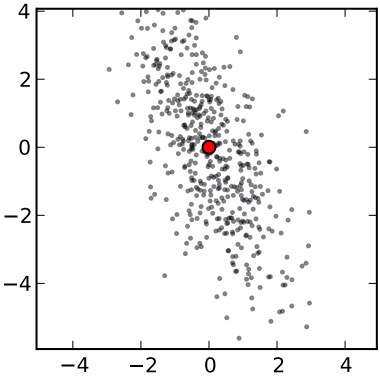
1. Find mean



Principal Component Analysis

universitätbonn

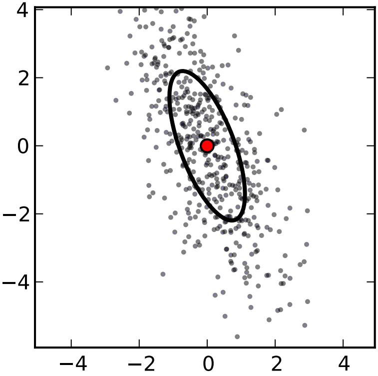
1. Find mean
2. Subtract mean



Principal Component Analysis

universitätbonn

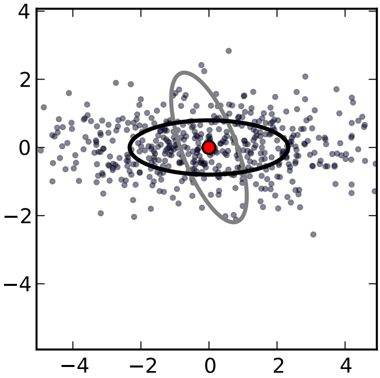
1. Find mean
2. Subtract mean
3. Model as ellipse



Principal Component Analysis

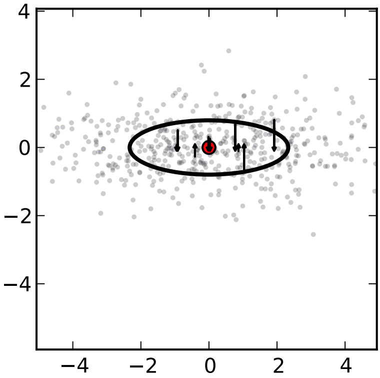
universitätbonn

1. Find mean
2. Subtract mean
3. Model as ellipse
4. Rotate to align with axis



Principal Component Analysis

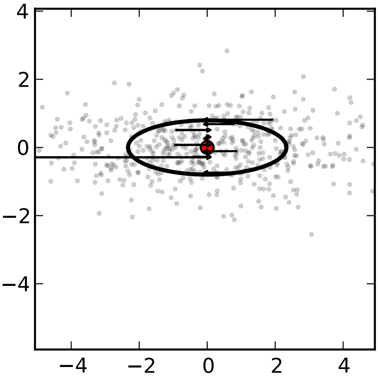
1. Find mean
2. Subtract mean
3. Model as ellipse
4. Rotate to align with axis
5. Project data points to 1st axis
   note the small error!



Principal Component Analysis

universitätbonn

1. Find mean
2. Subtract mean
3. Model as ellipse
4. Rotate to align with axis
5. Project data points to 1st axis
   note the small error!
6. Project data points to 2nd axis
   note the larger error!



Principal Component Analysis

universitätbonn

1. Find mean
2. Subtract mean
3. Model as ellipse
4. Rotate to align with axis
5. Project data points to 1st axis
   note the small error!
6. Project data points to 2nd axis
   note the larger error!
7. ...

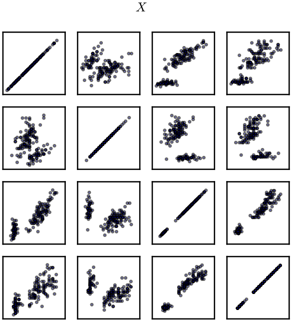
PCA Summary

- PCA projects to axis with greatest variance
- Often provides good first insight into dataset

$$\bar{X} \leftarrow X - \text{mean}(X) \qquad \bar{X} \in \mathbb{R}^{n \times N}$$

$$W \leftarrow \text{PCA}(\bar{X}, 2) \qquad W \in \mathbb{R}^{N \times M}$$

$$X_{\text{PCA}} \leftarrow \bar{X} \cdot W \qquad X_{\text{PCA}} \in \mathbb{R}^{n \times M}$$



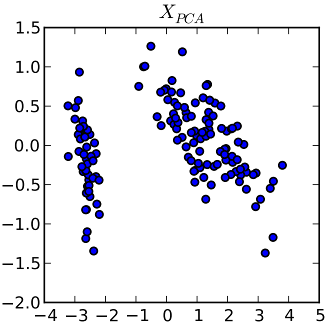
PCA Summary

- PCA projects to axis with greatest variance
- Often provides good first insight into dataset

$$\bar{X} \leftarrow X - \text{mean}(X) \qquad \bar{X} \in \mathbb{R}^{n \times N}$$

$$W \leftarrow \text{PCA}(\bar{X}, 2) \qquad W \in \mathbb{R}^{N \times M}$$

$$X_{\text{PCA}} \leftarrow \bar{X} \cdot W \qquad X_{\text{PCA}} \in \mathbb{R}^{n \times M}$$



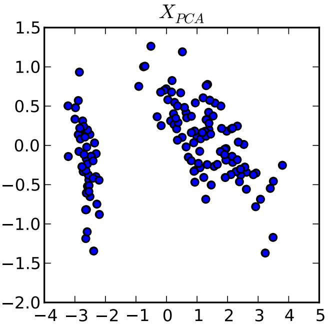
PCA Summary

- PCA projects to axis with greatest variance
- Often provides good first insight into dataset

$$\bar{X} \leftarrow X - \mathrm{mean}(X) \qquad \bar{X} \in \mathbb{R}^{n\times N}$$

$$W \leftarrow \mathrm{PCA}(\bar{X}, 2) \qquad W \in \mathbb{R}^{N\times M}$$

$$X_{\mathrm{PCA}} \leftarrow \bar{X} \cdot W \qquad X_{\mathrm{PCA}} \in \mathbb{R}^{n\times M}$$



- Identify important variables in projection matrix W:

```
W = [[ 0.36 -0.08 0.85 0.35]
     [-0.65 -0.72 0.17 0.07]]
```

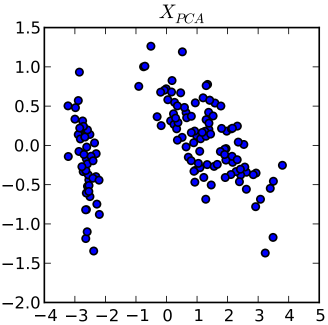
PCA Summary

- PCA projects to axis with greatest variance
- Often provides good first insight into dataset

$$\bar{X} \leftarrow X - \text{mean}(X) \qquad \bar{X} \in \mathbb{R}^{n \times N}$$

$$W \leftarrow \text{PCA}(\bar{X}, 2) \qquad W \in \mathbb{R}^{N \times M}$$

$$X_{\text{PCA}} \leftarrow \bar{X} \cdot W \qquad X_{\text{PCA}} \in \mathbb{R}^{n \times M}$$



- Identify important variables in projection matrix W:

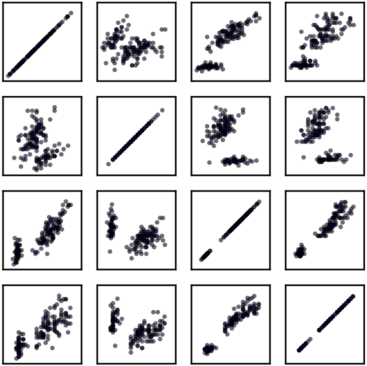
```
W = [[ 0.36 -0.08 0.85 0.35]
     [-0.65 -0.72 0.17 0.07]]
```

Noise Reduction

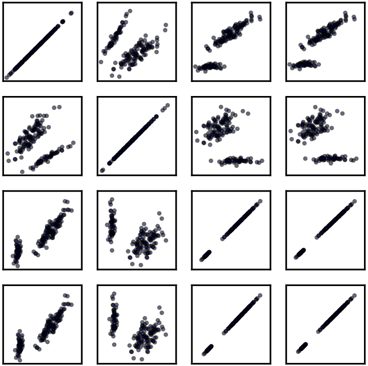
- Most of the data explained by first axes
- (almost) constant axes thrown away
- Projecting back to input-space reduces noise

$$X_{\text{clean}} \leftarrow X_{\text{PCA}} \cdot W^T + \text{mean}(X)$$

X



X_{clean}



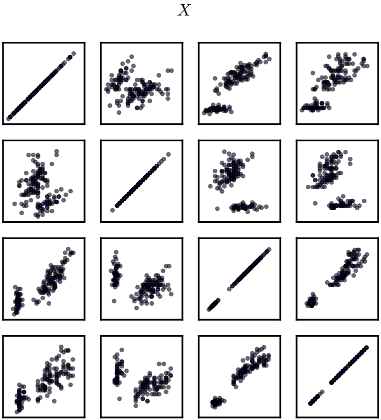
Interactive Part

- Open Notebook titled “1 - PCA”!

k-Means Motivation

universitätbonn

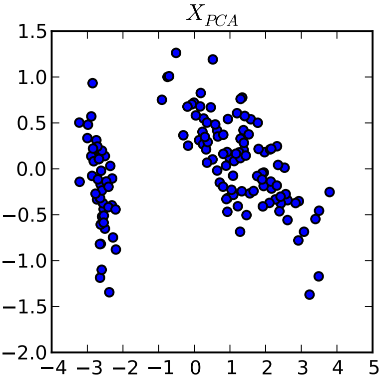
- Often, you don't have much information about the structure of X.
- In fact, we did not use any in the PCA step.



k-Means Motivation

universitätbonn

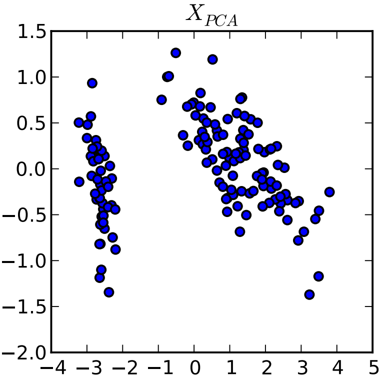
- Often, you don't have much information about the structure of X.
- In fact, we did not use any in the PCA step.
- By visualization, you can guess structure in X, "there might be 3 clusters".



k-Means Questions

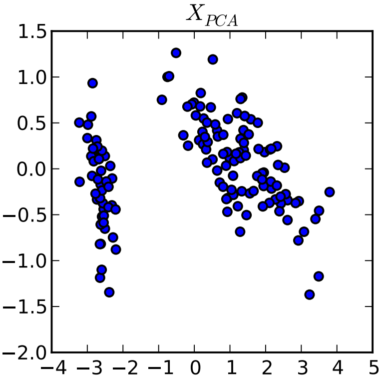
universitätbonn

1. Can we assign data points to clusters?



k-Means Questions

1. Can we assign data points to clusters?
2. Can we find a representative for each cluster?



k-Means Algorithm

k-Means finds assignments j and cluster centers μ by solving

$$\min_{\mu} \sum_{i=0}^{N} \min_{j} \left\| \mu_j - x_i \right\|^2 \tag{1}$$

The algorithm is simple:

1. Set μ, j to a random value
2. Solve (1) for j
3. Solve (1) for μ
4. If j or μ changed significantly, go to step 2.

k-Means Visualization

K-Means Website Example

Interactive Part

- Open Notebook titled “2 - KMeans”!

Outline

Supervised Learning – General

universitätbonn

- Task: Learn the function $y = f(x)$ which predicts the output y for the given input x, knowing the desired output
- Each example in data is a tuple of the input and desired output (target)

Example: Supervised Learning

universitätbonn

- Input Data: 40 examples of persons (age, height, smoker).
- Targets: Weight of the person (desired output)
- Goal: Learn a function which predicts the weight for the new person knowing the age, height, nationality of person.

Training / Test data

universitätbonn

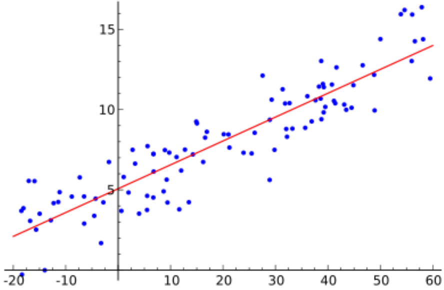
- Learning is done on the training data, for which we know the input and targets
- To test if the model learned to predict the output, we use test data.

Outline

Linear Regression

universitätbonn

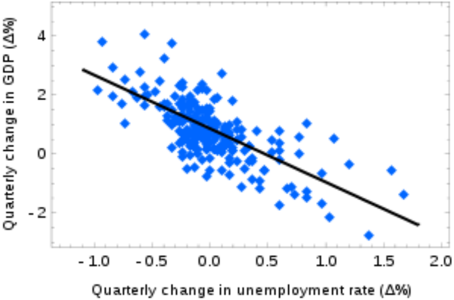
- Task: for the given input x predict the real value output $y = f(x)$
- Fit a hyperplane to data
- Linear function: simple, easy to understand.



Example: Okuns Law Quarterly Differences

universitätbonn

- Data: quarterly change in unemployment rate
- Predict: quarterly change in GDP



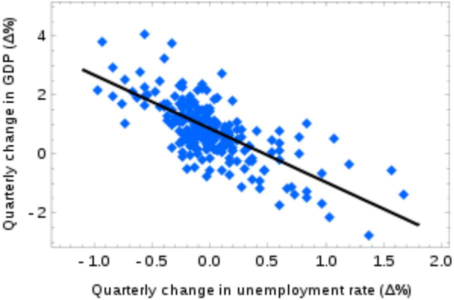
Mathematical Formulation I

universitätbonn

- Linear function: $y = \langle w, x \rangle + b$
- x - input vector
- w - weight vector
- b - bias
- y - output

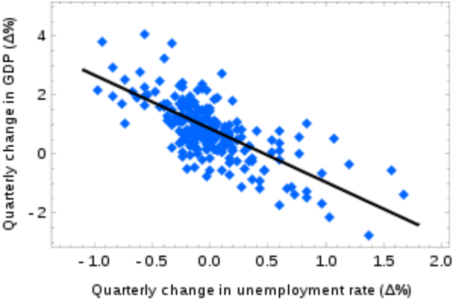
Example for Line

- $y = w_1 x_1 + b$
- How do we find coefficients w_i and bias b ?



Mathematical Formulation II

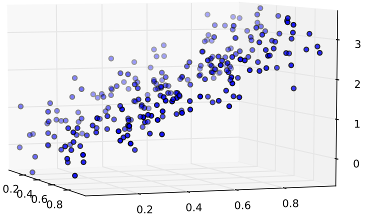
- Minimize the distance between each data point and the line
- $E = \sum_{i=0}^{N} (y_i - (w_i x_i + b))^2$
- Linear regression finds the weights and bias for which the error E is minimal



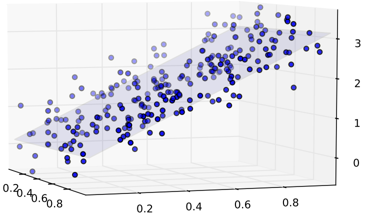
Example: 2D Data

universitätbonn

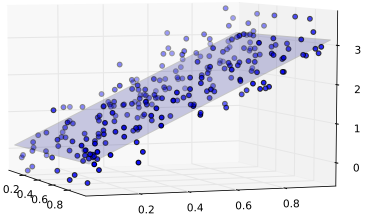
- What if our input data has 2-dim?
- We can see some linear relationship in the data



Example: 2D Data



- This time, we are fitting the plane
- $y = w_2 x_2 + w_1 x_1 + b$



Linear Regression – Interactive

universitätbonn

- Open Notebook titled 3a - Linear regression 1D.

Outline

Introduction to Python

Unsupervised Learning

Supervised Learning

- Linear Regression
- Classification
- Logistic Regression
- k Nearest Neighbors

Classification

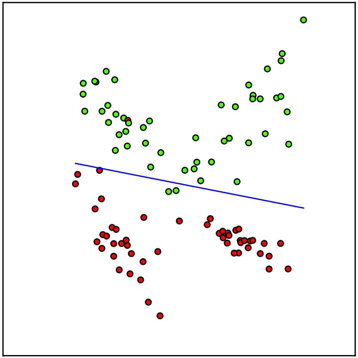
- Predict to which class a data point belongs.
- Training data are pairs $((x_0, y_0), \cdots, (x_N, y_N))$, $x_i \in \mathbb{R}^n, y_i \in \{0, \cdots, k\}$
- Classical example: Spam / Ham.
- All classes known beforehand.
- Other examples: Digit recognition, cancer benign/malignant, ...

Outline

Logistic Regression

universitätbonn

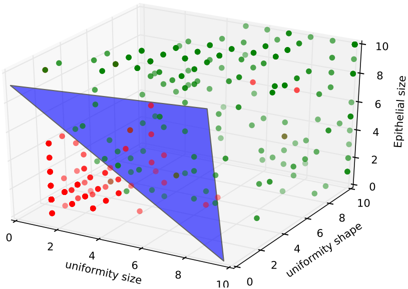
- Misnamed: Classification, not regression.
- Linear decision function: simple, easy to understand.



Example: Wisconsin Breast Cancer

universitätbonn

- Classify breast cancer samples in malign or benign.
- 700 Samples with 10 measurements each.
- We take only 3 measurements:
  - Uniformity of Cell Size
  - Uniformity of Cell Shape
  - Single Epithelial Cell Size
- Training on 525, test on 175
- 97.1% Accuracy

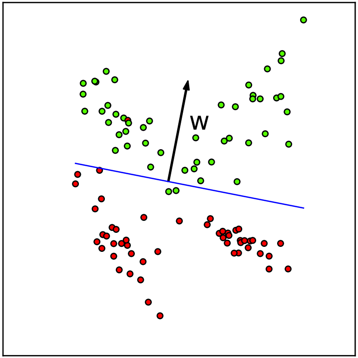


Mathematical Formulation I

universitätbonn

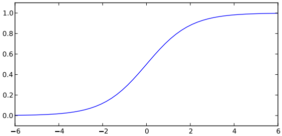
- For two classes $-1, +1$.
- Decision boundary given by hyperplane.
- Hyperplane defined by normal vector and offset:

$$y = \operatorname{sign}(\langle w, x \rangle + b)$$
$$w \in \mathbb{R}^n, b \in \mathbb{R}$$



Mathematical Formulation II

universitätbonn



- Relation to regression:

$$p(y = +1 \mid x) = \text{logistic}(\langle w, x \rangle + b)$$

- As probabilities are between 0 and 1, the logistic function squashes the regression result:

$$p(y = +1 \mid x) > 0.5 \Leftrightarrow \langle w, x \rangle + b > 0$$

- Need to solve:

$$\max_{w} \sum_{i=0}^{n} \log(p(Y = y_i | x_i))$$

Example: Classifying Insults I

universitätbonn

- Dataset: Forum posts / comments on social issues.
- Two classes: Insulting towards other posters / not insults.
- Training set: 4000 comments, test set: 2500 comments
- Features: Extract dictionary of all occuring words, count occurence per comment.
- Very high dimensional: 16.500

Either you are fake or extremely stupid...maybe both...

i really don't understand your point. It seems that you are mixing apples and oranges.

To engage in an intelligent debate with you is like debating to a retarded person. It's useless. It looks like you're bent on disregarding the efforts of the government.

@jdstorm dont wish him injury but it happened on its OWN and i DOUBT he's injured, he looked embarrassed to me

Example: Classifying Insults II

Either you are fake or extremely stupid...maybe both...

aaaah		are		feathers		olympic		stupid		you		zealot		zuckerberg
[0,	...,	1,	...,	0,	...,	0,	...,	1,	...,	1,	...,	0,	...,	0]

Example: Classifying Insults II

universitätbonn

Either you are fake or extremely stupid...maybe both...

aaaah		are		feathers		olympic		stupid		you		zealot		zuckerberg
[0,	...,	1,	...,	0,	...,	0,	...,	1,	...,	1,	...,	0,	...,	0]

Accuracy with logistic regression: 84.5%

Example: Classifying Insults II

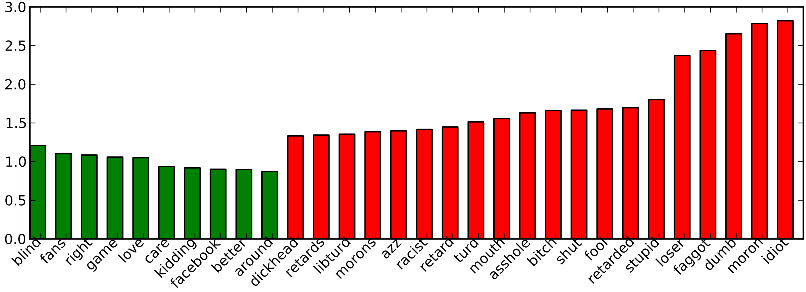
universitätbonn

Either you are fake or extremely stupid...maybe both...

aaaah		are		feathers		olympic		stupid		you		zealot		zuckerberg
[0,	...,	1,	...,	0,	...,	0,	...,	1,	...,	1,	...,	0,	...,	0]

Accuracy with logistic regression: 84.5%

The largest coefficients (sign given by color):



Interactive Part

universitätbonn

- Open Notebook titled “4 - Logistic Regression”.

Outline

Introduction to Python

Unsupervised Learning

Supervised Learning

- Linear Regression
- Classification
- Logistic Regression
- k Nearest Neighbors

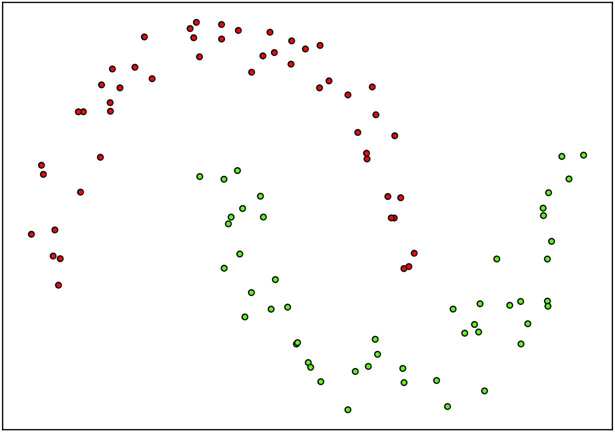
Nonlinear Problems

- Logistic regression works well if the data is linearly separable.
- Great for high dimensional data (such as text), not good for complicated low-dimensional data.

Nonlinear Problems

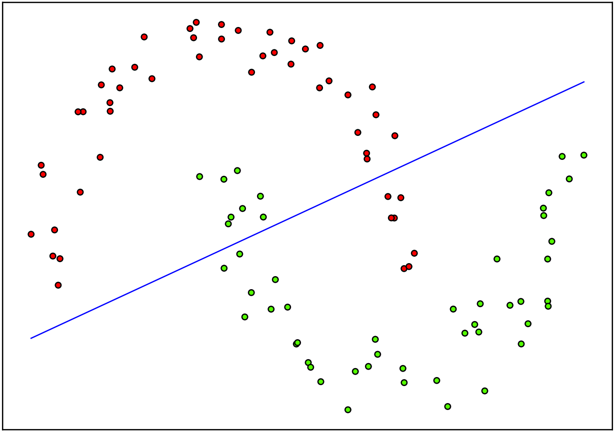
universitätbonn

- Logistic regression works well if the data is linearly separable.
- Great for high dimensional data (such as text), not good for complicated low-dimensional data.



Nonlinear Problems

- Logistic regression works well if the data is linearly separable.
- Great for high dimensional data (such as text), not good for complicated low-dimensional data.



k Nearest Neighbors

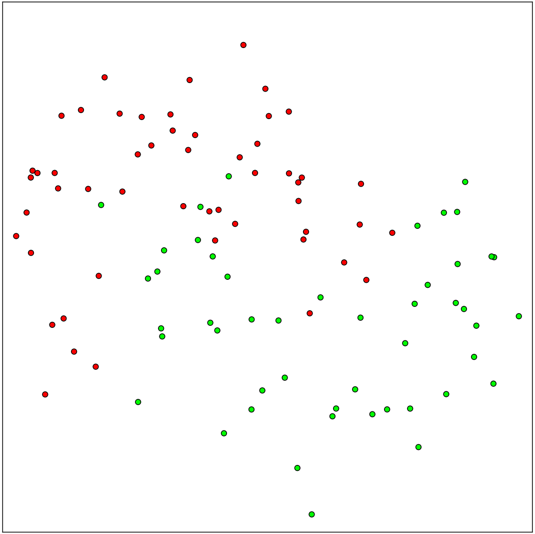
- Classification: same setup as logistic regression.
- Very simple but powerful idea: Do as your neighbors does.
- For a new point x look at the nearest (or the two nearest or three nearest, ...) point in the training data for a label.
- Usually: Euclidean distance in $\mathbb{R}^n$.

Simple algorithm

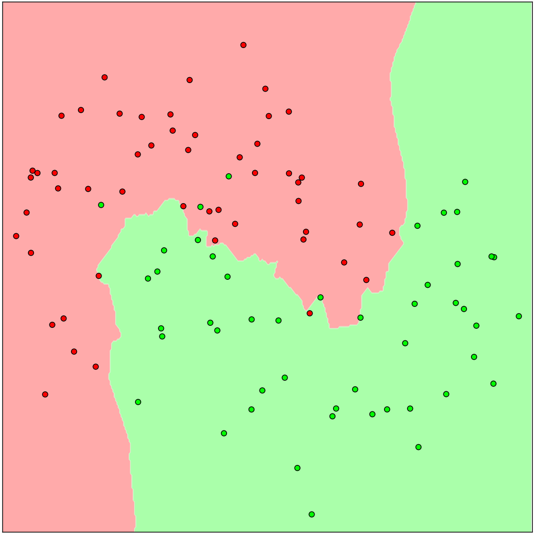
universitätbonn

- Pick a k, for example $k = 3$.
- Want to classify new example x.
- Compute $d_i = d(x_i, x)$, i.e. $d(x_i, x) = ||x_i - x||$.
- Sort d_i, take k smallest: $d_{i_0}, d_{i_1}, d_{i_2}$.
- Assign y that appears most often among $y_{i_0}, y_{i_1}, y_{i_2}$.

Illustration



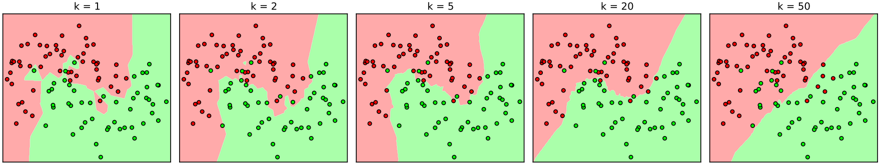
Illustration



$k = 5$

Picking *k*

- How do we choose *k*?
- General problem called model selection.
- For training data, $k = 1$ gives perfect prediction – but not for new data!



Picking k (in practice)

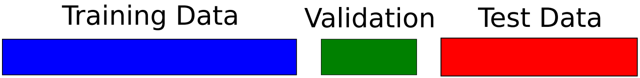
- We can not choose k on the training set.
- We can not choose k on the set we evaluate our algorithm on (or for that we need predictions).

Picking k (in practice)

- We can not choose k on the training set.
- We can not choose k on the set we evaluate our algorithm on (or for that we need predictions).

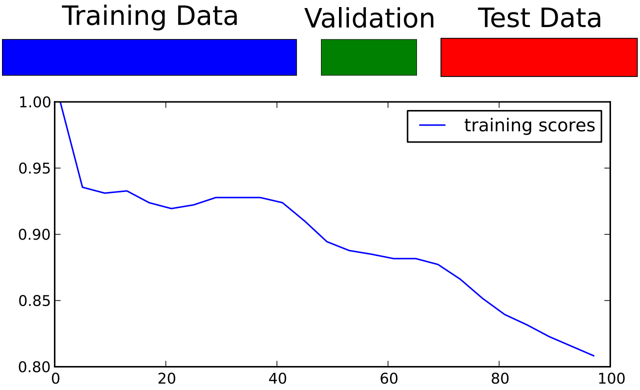
Picking k (in practice)

- We can not choose k on the training set.
- We can not choose k on the set we evaluate our algorithm on (or for that we need predictions).



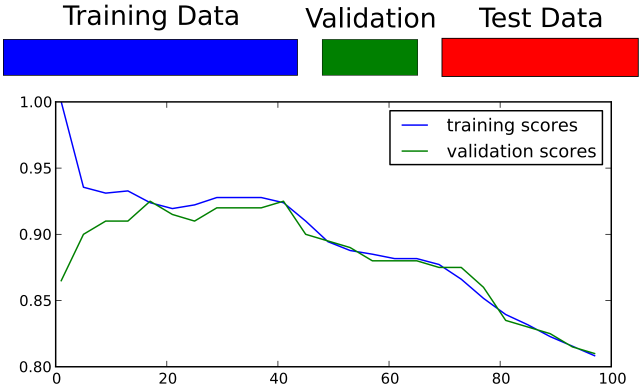
Picking k (in practice)

- We can not choose k on the training set.
- We can not choose k on the set we evaluate our algorithm on (or for that we need predictions).



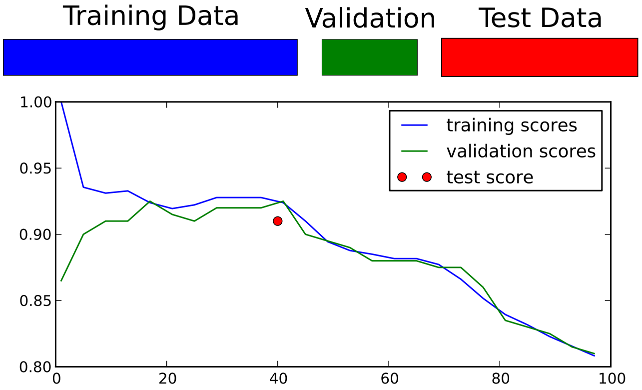
Picking k (in practice)

- We can not choose k on the training set.
- We can not choose k on the set we evaluate our algorithm on (or for that we need predictions).



Picking *k* (in practice)

- We can not choose *k* on the training set.
- We can not choose *k* on the set we evaluate our algorithm on (or for that we need predictions).



Interactive Part

- Open Notebook titled “5 - k Nearest Neighbors”.