pawel.kasprowski@polsl.pl

Deep Learning in Python

Introduction to Data Mining

Paweł Kasprowski, PhD, DSc.

Problems to be solved

- Classification
  - assign samples to predefined classes
  - [photo of an animal -> species]
  - [text -> author]
  - [scan-path -> novice/expert]
- Regression
  - calculate a value for each sample
  - [images of the house -> price of the house]
  - [eye image -> gaze coordinates]
- Conversion
  - convert object into another object
  - [image -> text description of the image]
  - [voice recording -> text of the speech]

Task and method

- Task: find a function Y = f(X) where
  - X – **sample** (input)
  - Y – result – class, value (output) – **label**
- Method (learning by example):
  - take some number of samples X with known output Y (**examples**, labeled samples)
  - build the function based on these examples (learning process)
  - use the function to predict the label for unknown samples

Problem of "generality"

- It is relatively easy to create a function that correctly classifies all examples
- But is this function "general" enough?
  - is it able to correctly classify unknown samples?
- The answer is not trivial and that is why we have a lot of different classification methods
- "No free lunch" theorem

Over-fitting

- If the function (model) is optimized for the given data (given examples) it may have poor generality
- This problem is called over-fitting
- Solution is to simplify the model, e.g. stop learning even when it is not perfect for examples

The simplest algorithms

- K-Nearest Neighbors (kNN)
  - a new sample is classified as majority of its K nearest neighboring examples
- Linear Discriminant Analysis (LDA)
  - finds a linear equation separating positive and negative samples
- Naive Bayes
  - the class of a new sample is calculated based on examples using Bayes equation
- Decision Tree
  - the tree is built using examples by seaching for the most discriminative features
  - new samples are going through the tree reaching a leaf with the predicted class

More sophisticated algorithms

- Random Forest
  - a set of randomly generated decision trees voting for the final results
- Support Vector Machines (SVM)
  - recalculates examples into another space where they are easier to separate lineary (kernel trick)
- Neural Networks
  - based on neurons organized into layers

Hyperparameters

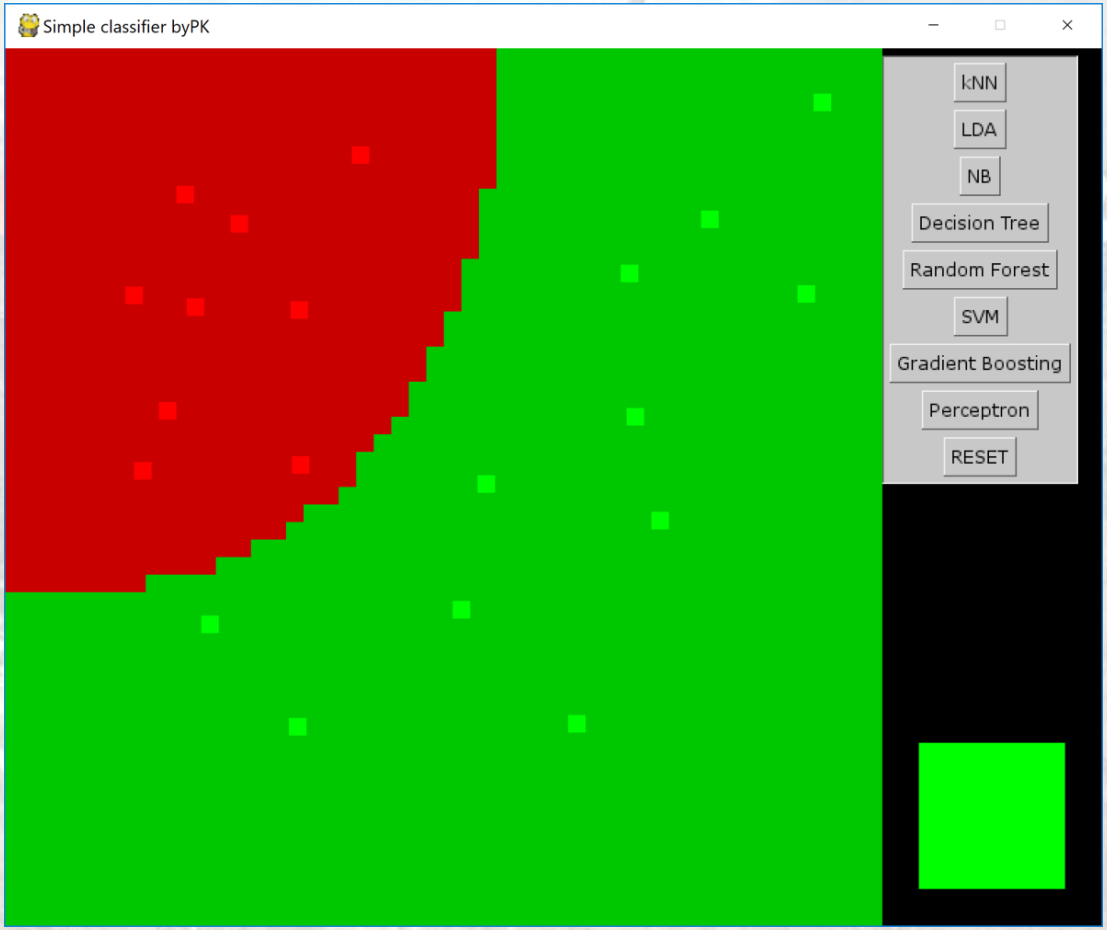
- Classification algorithms have parameters
  - (so called hyperparameters)
- kNN – number of neighbors
- Random Forest – number of trees
- SVM – kernel function, parameters of the function, learning rate etc.
- Good choice of hyper parameters influences the quality of the model!
- Bad news: deep learning methods have A LOT hyper parameters

Scikit-learn

- Package implementing:
  - many popular classification algorithms
  - many classic methods for data handling
- General use:
  - load the dataset consisting of samples and their labels
  - create a model using one of the possible algorithms (classes)
  - train the model: model.fit(samples, labels)
  - use the model to predict classes:
    - predicted = model.predict(sample)

Universal example

python classifier



Creating a model

```
def classification(model_name,samples,labels,rangex,rangey):
models = {
        "KNN": KNeighborsClassifier(),
        "LDA": LinearDiscriminantAnalysis(),
        "NB": GaussianNB(),
        "TREE":DecisionTreeClassifier(),
        "RF":RandomForestClassifier(n_estimators=20),
        "SVM":SVC(gamma='scale'),
        "PERC":Perceptron(max_iter=2000),
        "GB":GradientBoostingClassifier()
    }
model = models.get(model_name)
...
```

Using the model

```
...
samples = np.array(samples)
labels = np.array(labels)
model.fit(samples, labels)
# build the matrix of results using the model
result = np.zeros([rangex,rangey])
for x in range(rangex):
    for y in range(rangey):
        sample = np.array([x,y])
        result[x][y] = model.predict(sample.reshape(1, -1))
return result
```

Classification of a single sample

- result = model.predict(samples)
  - samples must have the dimension:
    - [number of samples, number of attributes]
  - For one sample: [1,2]
- Code version 1:
  - sample = np.array([x,y]) # dimension [2,]
  - result[x][y] = model.predict(sample.reshape(1, -1))
- Code version 2:
  - sample = np.array([[x,y]]) # double brackets - dimension [1,2]
  - result[x][y] = model.predict(sample)

Classification of all samples

- It is possible to classify all samples at once
- Create a list of points to classify:

```
samples = []
for x in range(rangex):
    for y in range(rangey):
        samples.append((x,y))
```

- A more pythonic way:

```
samples = [(x,y) for x in range(rangex) for y in range(rangey)]
```

- Convert to numpy

```
samples = np.array(samples)
```

Classification of all samples at once

- Predictions for all points

```
r = model.predict(samples)
```

- Create the array of results (shape: [rangex,rangey])

```
result = np.zeros([rangex,rangey])
for i in range(len(samples)):
    result[samples[i,0],samples[i,1]]=r[i]
```

- Is it possible to do it without the "for" loop?

Example for the iris dataset

iris.ipynb

```
import pandas as pd
from sklearn.neighbors import KNeighborsClassifier
data = pd.read_csv('iris.data')

model = KNeighborsClassifier()
samples = data.values[:,0:4] # first four columns
labels = data.values[:,4] # the fifth column
model.fit(samples, labels)
predicted = model.predict([(5.9, 3.0, 5.1, 1.8)])
print("predicted",predicted)
> predicted ['Iris-virginica']
```

	sl	sw	pl	pw	iris
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Evaluating results

- Hyperparameters may be tuned
  - it is important to be able to check if the model is general
- First rule: never check the model using the same data that you used for training (examples)
  - over-fitting!
- The dataset should always be divided into:
  - training set (used to build the model)
  - test set (to check how it works for unknown samples)

Evaluation

- Dividing into training set and test set may be not enough!
  - we optimize the model (by tuning hyper parameters) for the given test set
- The safest way:
  - training set (for learning)
  - validation set (for hyper parameters tuning)
  - test set (for final evaluation)
    - the "holdout" dataset

Sklearn train_test_split

- (trainSamples, testSamples, trainLabels, testLabels) = sklearn.model_selection.train_test_split(samples, labels)
- Two pairs:
  - trainSamples, trainLabels
  - testSamples, testLabels
- Default: 75% train, 25% test
- Parameters:
  - train_size – percent or number of samples
  - test_size – percent or number of samples

iris.ipynb / Train Test Split

Stratification

- Problem: proportion of samples for classes may differ in training and test sets
- For instance: 50-50-50 in the original set
  - training: 40-30-25
  - test: 10-20-25
- Effect: class0 will be treated as more probable!
- Solution: stratified division (maintain distribution)
- Parameter: stratify=labels

Check test samples

iris.ipynb / Check an error

- Checking the test set

```
correct = 0;
predictedLabels = model.predict(testSamples)
for i in range(len(testSamples)):
    print(testLabels[i],"->",predictedLabels[i],end=' ')
    if(testLabels[i]==predictedLabels[i]):
        correct = correct + 1; print('OK')
    else:
        print('error!!!')
print("Correct: {} of {} accuracy = {:.2f}"
      .format(correct,len(testSamples),correct/len(testSamples)))
```

Cross-validation

- Dynamic division into training and test
  - the same examples are sometimes training and sometimes test samples
- 10-fold cross validation:
  - divide the whole dataset into 10 subsets
  - for each subset
    - train the model using the remaining 9 subsets
    - test the results using the chosen subset
  - average the results
- Folds are randomized but may be stratified
  - the same distribution of classes in each fold

Sklearn cross_validate

- ***iris.ipynb / Cross validation***
  - sklearn.model_selection.cross_validate(model, samples, labels, cv=<number of folds>)
- Returns for each fold:
  - fit_time
  - score_time
  - test_score
- Many parameters to check other metrics!

Measures

- Accuracy – the number of correctly classified samples to all samples
  - a problem with unbalanced sets
  - if 90% of test samples is positive the blind classifier achieves 90% accuracy
- Precision – the cost of false positives
  - not belonging to class predicted as belonging
- Recall – the cost of false negatives
  - belonging to class predicted as not belonging
- F1-Score – combination of precision and recall
- Cohen's kappa – compares prediction with random prediction

Measures interpretation

- Example: We have built the model that diagnoses COVID-19 based on a genetic test
- For every eye movement sample the model returns:
  - ***H*** - if a person is healthy
  - ***S*** - when a person is sick
- We test the model using
  - 18 samples from sick people (***S***)
  - 82 samples from healthy people (***H***)
- We achieve 82% accuracy
  - is it bad or good?
  - it depends: we must examine a confusion matrix

Confusion matrix

Predicted as >> Real class	H	S
H	81	1
S	17	1

P	N
TP	FN
FP	TN

Accuracy 0.82

H:
precision: 0.83
recall: 0.99
F1-score: 0.90

S:
precision: 0.50
recall: 0.06
F1-score: 0.1

Cohen's kappa: 0.07

Measures

Predicted as >> Real class	H	S
H	80	2
S	15	3

P	N
TP	FN
FP	TN

Accuracy 0.83

H:
precision: 0.84
recall: 0.98
F1-score: 0.90

S:
precision: 0.60
recall: 0.17
F1-score: 0.26

Cohen's kappa: 0.20

Measures

Predicted as >> Real class	H	S
H	66	16
S	2	16

P	N
TP	FN
FP	TN

Accuracy 0.82

H:
precision: 0.97
recall: 0.80
F1-score: 0.88

S:
precision: 0.50
recall: 0.89
F1-score: 0.64

Cohen's kappa: 0.53

Measures

Predicted as >> Real class	H	S
H	64	18
S	0	18

P	N
TP	FN
FP	TN

Accuracy 0.82

H:
precision: 1.00
recall: 0.78
F1-score: 0.88

S:
precision: 0.50
recall: 1.00
F1-score: 0.67

Cohen's kappa: 0.56

Example for three classes

Confusion matrix

```
[[24  3  9]
 [ 3 24 24]
 [ 2 15 49]]
```

class	precision	recall	f1-score	support
0	0.83	0.67	0.74	36
1	0.57	0.47	0.52	51
2	0.60	0.74	0.66	66
micro avg	0.63	0.63	0.63	153
macro avg	0.67	0.63	0.64	153
weighted avg	0.64	0.63	0.63	153

Accuracy: 0.63

Scikit learn measures

- Useful functions from *sklearn.metrics*:
  - confusion_matrix(labels, results)
  - classification_report(labels, results)
  - accuracy_score(labels, results)
  - precision_score(labels, results)
  - recall_score(labels, results)
  - cohen_kappa_score(labels, results)
  - ...

Measures calculation

iris.ipynb / Measures calculation

```
from sklearn.metrics import classification_report, confusion_matrix,
    accuracy_score,cohen_kappa_score


predictedLabels = model.predict(testSamples)
print(confusion_matrix(testLabels, predictedLabels))
print(classification_report(testLabels, predictedLabels))
accuracy = accuracy_score(testLabels, predictedLabels)
print("Accuracy: {:.2f}".format(accuracy))
c_kappa = cohen_kappa_score(testLabels, predictedLabels)
print("Cohen's Kappa: {:.2f}".format(c_kappa))
```

Adding own scorers to cross validation

iris.ipynb / Adding own scorers to cross validation

```
from sklearn.metrics import make_scorer, precision_score, recall_score

p_scorer = make_scorer(precision_score, average='micro')
r_scorer = make_scorer(recall_score, average='micro')
my_scorer = {'precision': p_scorer, 'recall': r_scorer}

sklearn.model_selection.cross_validate(model, samples, labels, cv=5,
    scoring=my_scorer)
```

Summary

- Accuracy is often not enough
  - especially for unbalanced datasets
- Cohen's kappa is the reliable way to evaluate results
  - but has some drawbacks as well!
- The evaluation depends on the target we want to achieve
  - False Acceptance Rate (FAR=0)
    - no incorrectly classified healthy people
  - False Rejection Rate (FRR=0)
    - no incorrectly classified sick people

Regression

- The model does not search for one of N classes but for a value
  - E.g. price of the house
- Examples:
  - Linear Regression
  - Support Vector Regression
  - Decision Tree Regression
  - Neural Network Regression

Evaluation of regression

- Mean Absolute Error
- Mean Squared Error
  - penalizes predictions differing greatly (outliers)
- Root Mean Squared Error
  - comparable to real values
- Coefficient of Determination (R^2)
  - value 0-1
  - negative value if worse than random

Example: wine quality dataset

- Every wine has 11 numerical features
  - 1 - fixed acidity
  - 2 - volatile acidity
  - 3 - citric acid
  - 4 - residual sugar
  - 5 - chlorides
  - 6 - free sulfur dioxide
  - 7 - total sulfur dioxide
  - 8 - density
  - 9 - pH
  - 10 - sulphates
  - 11 - alcohol

- The result: wine quality in 0-10 scale
- Dataset: 4898 wines

Wine dataset analysis

wine.ipynb

- Simple classification
- Train-test split
- Measures calculation – confusion matrix
- Problem:
  - quality is an ordinal value not just a class
  - results should be calculated as a distance from the correct one
    - it is not bad when wine with quality 3 is classified as 4
    - it is a big problem when it is classified as 9

Regression measures

- Errors:

```
from sklearn.metrics import mean_squared_error, r2_score
print('MSE=',mean_squared_error(testLabels, predictedLabels))
print('R2=',r2_score(testLabels, predictedLabels))
```

- Results:
  - MSE = 0.99
  - R2 = -0.27

Linear Regression

- Using Linear Regression:

```
from sklearn import linear_model
model = linear_model.LinearRegression()
model.fit(trainSamples, trainLabels)
predictedLabels = model.predict(testSamples)
```

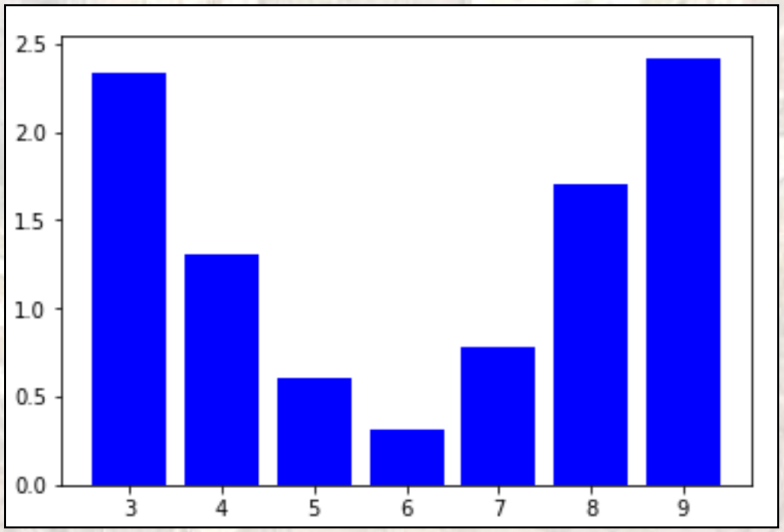
- The result for each sample is a float!
- Errors:
  - MSE = 0.53
  - R2 = 0.31

Errors per class estimation

- Mean errors per class

```
errors = np.abs(modelResults-testLabels)
for i in range(3,10):
    print('class',i,' avg error=',errors[np.where(testLabels==i)].mean())
```

- Results
  - class 3 avg error= 2.33
  - class 4 avg error= 1.31
  - class 5 avg error= 0.61
  - class 6 avg error= 0.31
  - class 7 avg error= 0.78
  - class 8 avg error= 1.71
  - class 9 avg error= 2.42



Problem: inbalanced classes

- Some qualities are rare, some are common
- Solution: Calculate weight for each class:

```
from sklearn.utils import class_weight
class_weights = class_weight.compute_class_weight('balanced',
                                                  classes=np.unique(trainLabels),y=trainLabels)
weights = np.ones([len(trainLabels)]) # initialize all weights to 1
for i, label in enumerate(trainLabels):
    weights[i] *= class_weights[int(label-3)] # the first label is 3!
```

- Using weights:

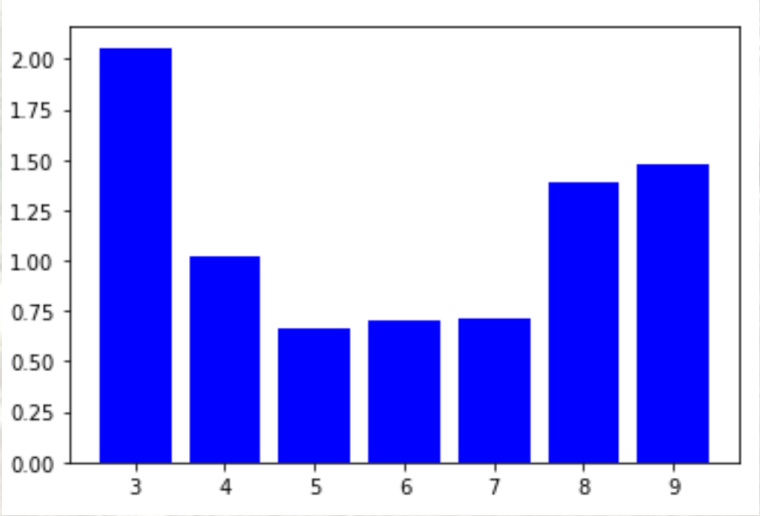
```
model.fit(trainSamples, trainLabels, sample_weight=weights)
```

Regression with weights

- Mean errors per class

```
errors = np.abs(modelResults-testLabels)
for i in range(3,10):
    print('class',i,' avg error=',errors[np.where(testLabels==i)].mean())
```

- Results
  - class 3 avg error= 2.06
  - class 4 avg error= 1.02
  - class 5 avg error= 0.66
  - class 6 avg error= 0.70
  - class 7 avg error= 0.72
  - class 8 avg error= 1.39
  - class 9 avg error= 1.48



Results

- Distribution is more equal for the weighted model
- But the error is higher (both MSE and R2)
- Question:
  - what do we really need?
  - what is our objective?
- Possibilities:
  - minimize absolute error
  - minimize MSE/RMSE
  - minimize error for high quality wines
  - …

Feature selection

- Not every feature is valuable
  - adds some knowlegde about a class
  - e.g. bottle's color is *probably* not significant for wine quality
- Not important features may spoil classification!
- There are plethora of algoritms that aim at selecting only the relevant features from the dataset
  - topic for a new lecture (or even a new subject)
- Sklearn has a very convenient class to solve the problem

SelectKBest

- Code:

```
from sklearn.feature_selection.univariate_selection import
    SelectKBest

newSamples = SelectKBest(<algorithm>, k=<number>)
    .fit_transform(samples, labels)
```

- Available algorithms:

(see: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

f_classif, mutual_info_classif, chi2, f_regression, mutual_info_regression, SelectPercentile, SelectFpr, SelectFdr, SelectFwe, GenericUnivariateSelect

Example

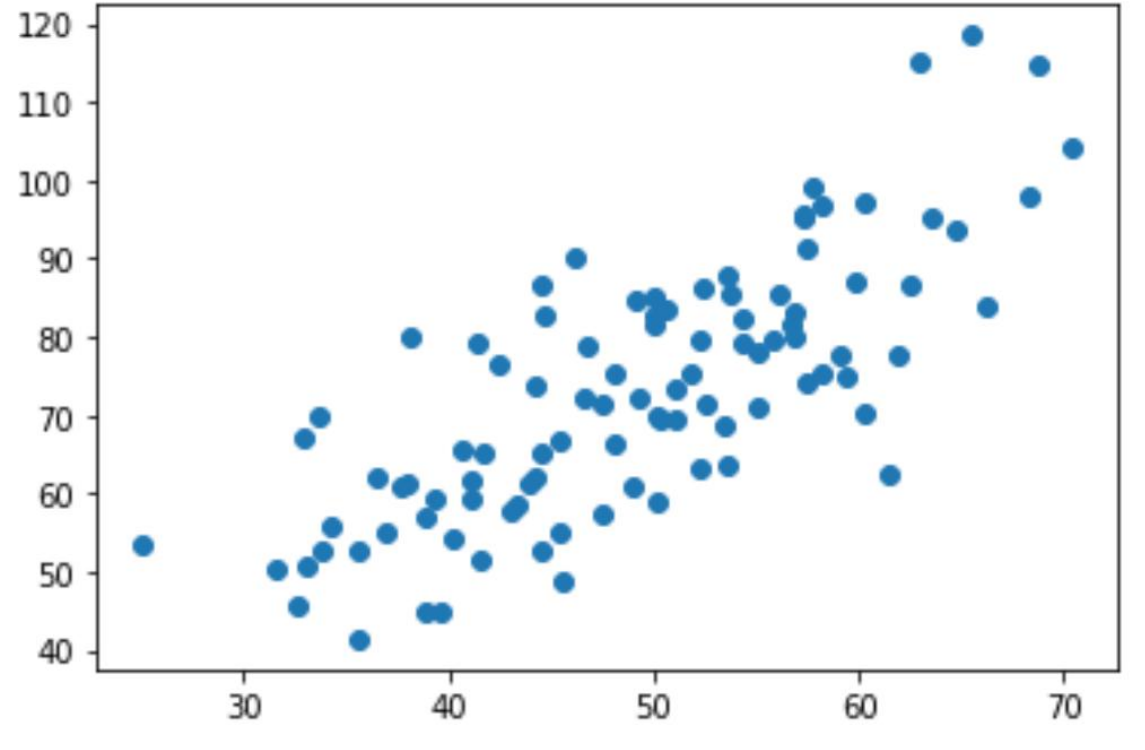
- Select 5 best features using f_regression algorithm:

```
from sklearn.feature_selection.univariate_selection import SelectKBest
print("Samples before",samples.shape)
newSamples = SelectKBest(sklearn.feature_selection.f_regression,k=5)
    .fit_transform(samples, labels)
print("Samples after", newSamples.shape)
samples = newSamples
```

- Output:
  - Samples before (4898, 10)
  - Samples after (4898, 5)
- Task for you: check if it works better!

Simple linear regression

- Find the best linear approximation of points



Example: Linear Regression

- Given a training set of pairs (X,Y) find coefficients of a linear function

$$y = a * x + b$$

- Error (loss) function Mean Squared Error

$$E = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i}^{n}(y_i - (ax_i + b))^2$$

- Goal: minimize this function

Linear Regression using Gradient Descent

- Gradient Descent algorithm:

  initialize randomly a and b

  repeat:

    calculate error (loss)

    change a and b in such way that error is smaller

  until error is small enough or it is not decreasing

- Question: how to find the direction of change for ***a*** and ***b***?

Gradient Descent

- Gradient dE/da – a function describing how ***E*** changes when ***a*** changes

$$\frac{\delta E}{\delta a} = \frac{-2}{n}\sum_{i}^{n}(\hat{y}_i - y_i)x_i$$

- Gradient dE/db - function describing how ***E*** changes when ***b*** changes

$$\frac{\delta E}{\delta b} = \frac{-2}{n}\sum_{i}^{n}(\hat{y}_i - y_i)$$

Gradient Descent algorithm

- for each step (iteration)
  - calculate error E
  - calculate gradients dE/da and dE/db
  - a = a - learning rate * dE/da
  - b = b - learning rate * dE/db
- Every iteration gives better values of a and b
  - error (**loss**) is smaller
- Learning rate controls how fast the parameters are updated
  - too fast – we may pass the minimum
  - too slow – we may wait forever

Analytical solution

- For linear regression the best parameters may be calculated using equations:

$$a = \frac{\sum_i^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^N (x_i - \bar{x})^2}$$

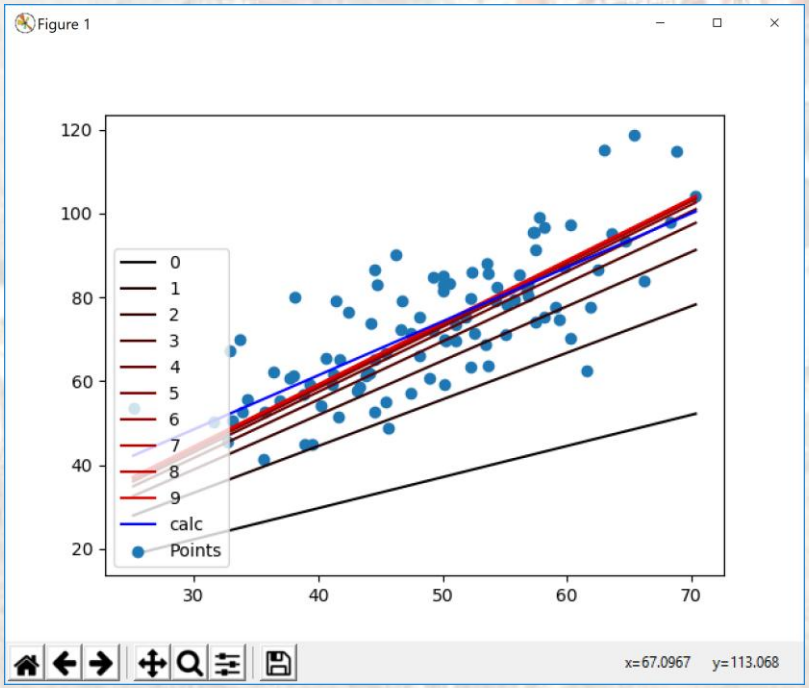
$$b = \bar{y} - a\bar{x}$$

- but the presented example shows how Gradient Descent algorithm works (and we will use this algorithm later!)

Example

linear_descent.ipynb

- Simple code calculating linear regression using gradient descent for a set of points

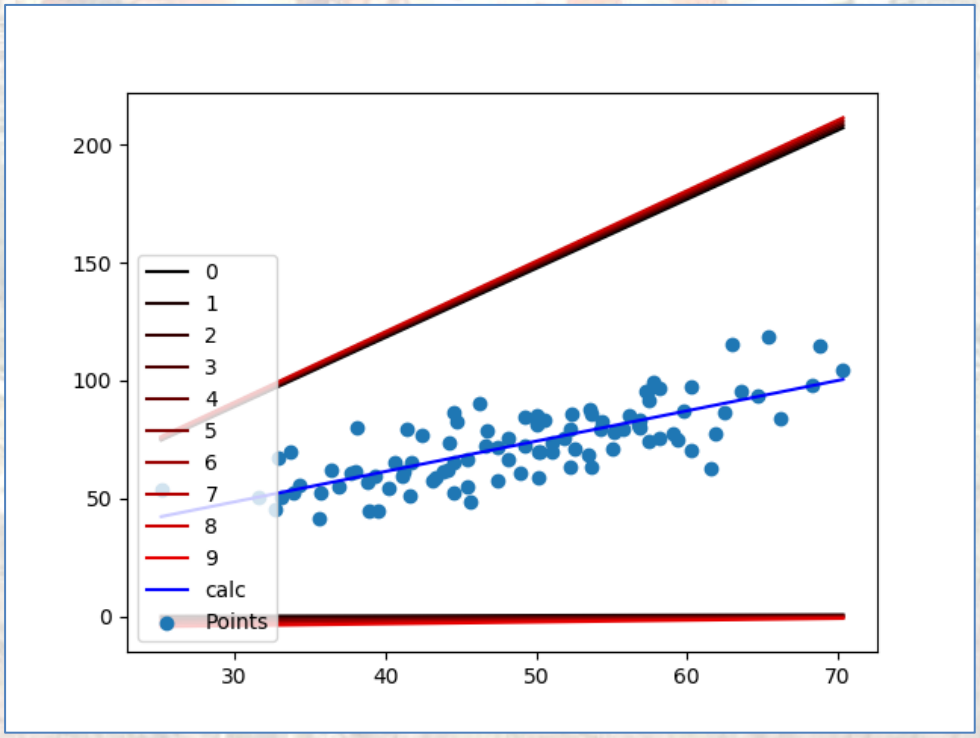


Observations

- If we change initial values of a and b we have different results
  - b is reluctant to change
  - kind of "vanishing gradient" problem
- (1) Change the learning rate for b:
  - nb = b - L*100 * D_b # Update b
- (2) Change the number of iterations
  - epochs = 100 (1000?)
- The results are better!

Learning rate

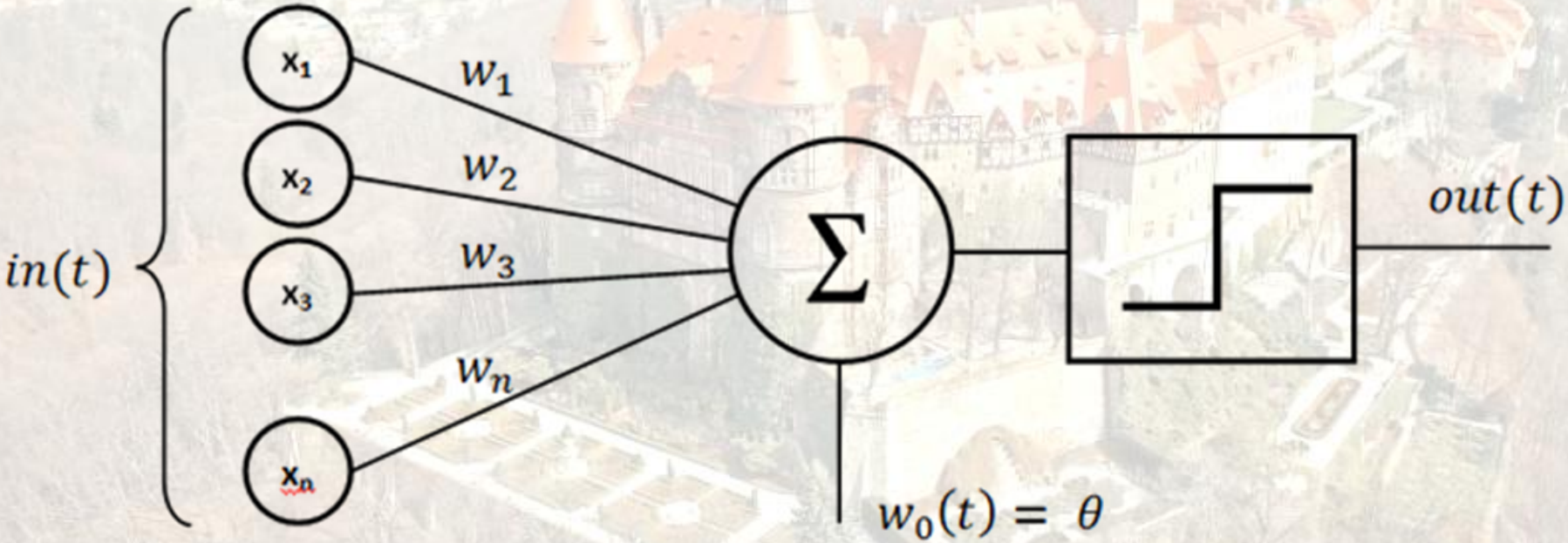
- Change to:
  - L = 0.0004 # The learning rate
- Results jump over the solution
  - too big step!



Summary

- There are many different classification algorithms
- There are many ways to tune these algorithms
- There are many measures to estimate the quality of classification/regression
- So what is so special about the deep learning methods?
  - Wait for the following sections!

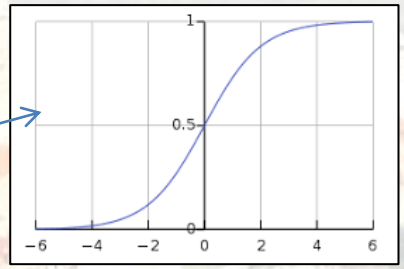
Perceptron (Rosenblatt 1957)

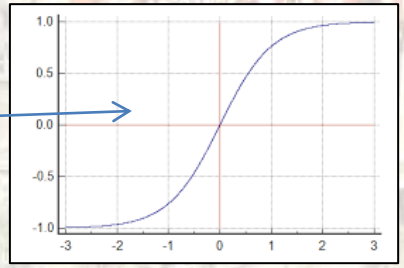


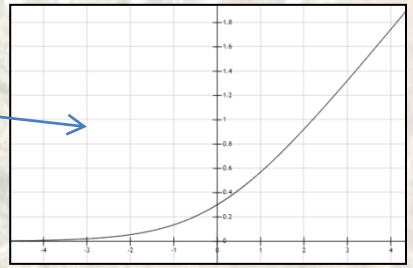
https://pl.wikipedia.org/wiki/Perceptron

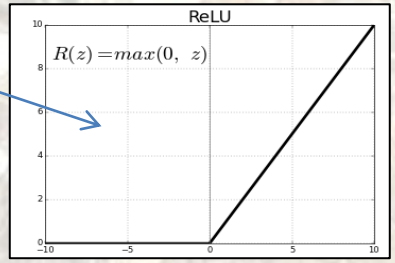
Activation function

- Linear combination of inputs
- Activation functions may be different
  - threshold function
  - sigmoid function
  - tanh function
  - softplus
  - RELU



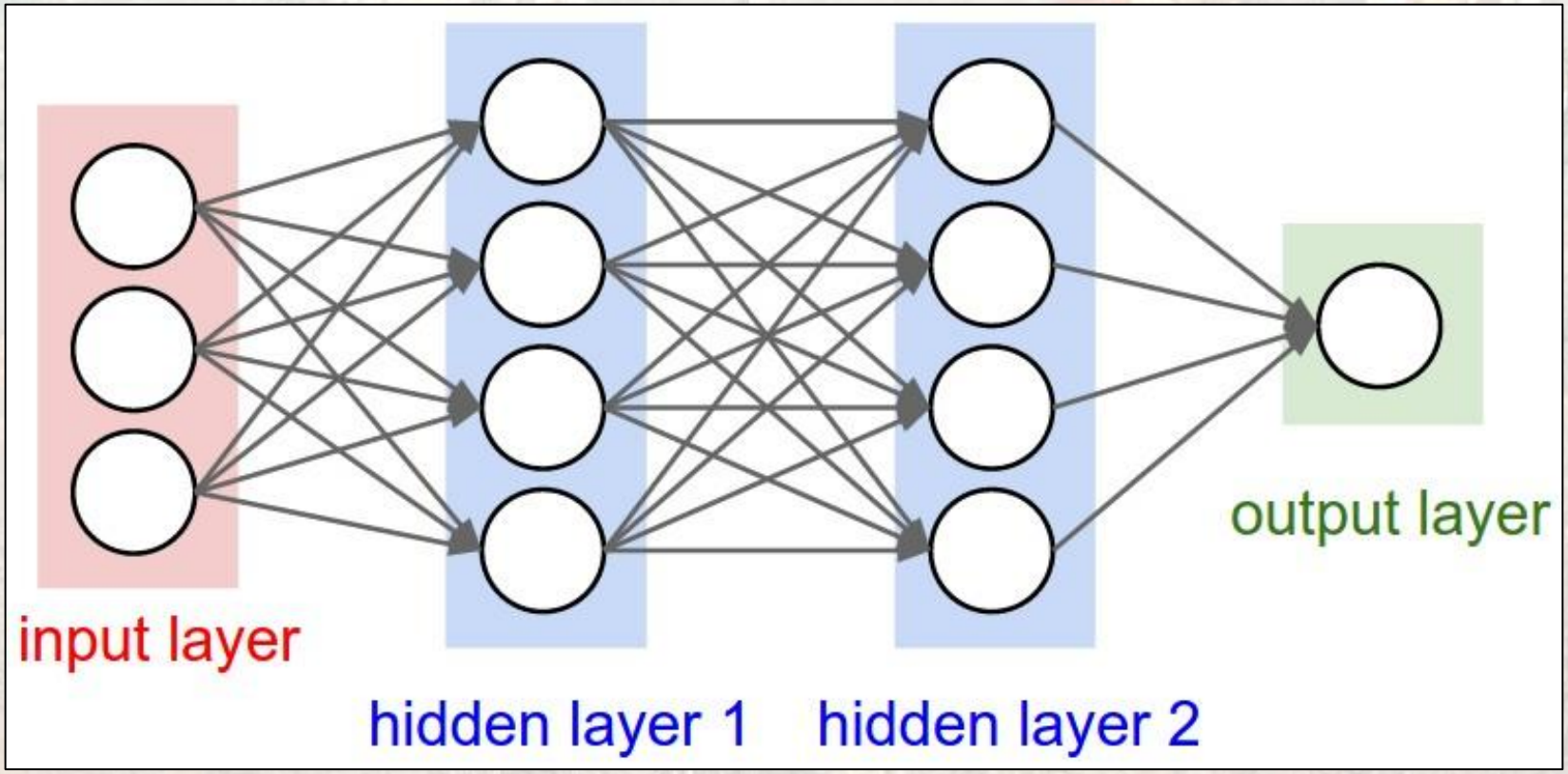






Artifical Neural Network (ANN)

- input layer > hidden layers > output layer



http://cs231n.github.io/convolutional-networks/

Training the network

- Back propagation
  - the state-of-the art method to train the network
  - utilizes the already mentioned Gradient Descent algorithm
- The algorithm:

  initialize weights

  repeat many times:

    use network to calculate output (Y_{pred}) for some examples (X)

    calculate error (loss function) using Y_{pred} and Y (real)

    update weights to minimize loss (using gradient for direction)

Tuning - hyperparameters

- Network structure
  - Number of layers
  - Number of neurons for layer
  - Connections
  - Activation functions for layers
- Loss function
- Optimization algorithm
  - how to change the weights
  - learning rate (how big changes of weights)

Implementations

- Many classification libraries like scikit-learn or WEKA implement the neural networks
  - but only the simplest models
- For Deep Learning there are many libraries developed by leading companies:
  - Tensorflow, Google
  - PyTorch, Facebook
  - CNTK, Microsoft
  - …
- We will use the most popular: Keras/Tensorflow

Keras

- General interface to use deep networks
- Works with Tensorflow, Theano, CNTK...
- User-friendly, modular, and extensible
- Created in Google
- Works in Python
- Tensorflow library includes Keras by default
  - that is why we installed Tensorflow and we use Keras
- Tensorflow 2 is integrated with Keras

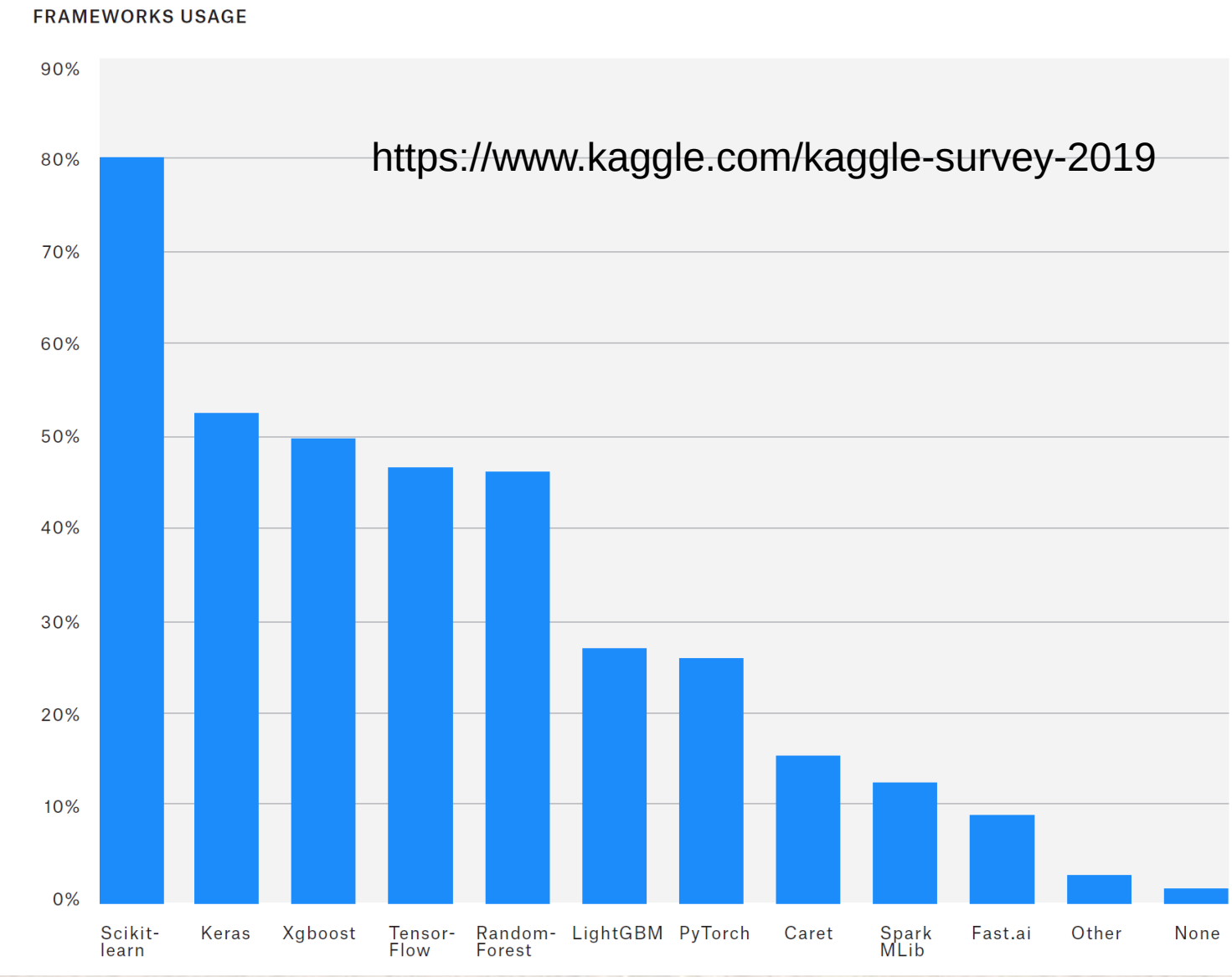

FRAMEWORKS USAGE
https://www.kaggle.com/kaggle-survey-2019
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Scikit-learn
Keras
Xgboost
Tensor-Flow
Random-Forest
LightGBM
PyTorch
Caret
Spark MLib
Fast.ai
Other
None

Table 3: Major Deep Learning Platforms

Platform	2019 % share	2018 % share	% change
Tensorflow	31.7%	29.9%	5.8%
Keras	26.6%	22.2%	19.7%
PyTorch	11.3%	6.4%	75.5%
Other Deep Learning Tools	5.6%	4.9%	15.2%
DeepLearning4J	2.5%	3.4%	-25.6%
Apache MXnet	1.7%	1.5%	13.1%
Microsoft Cognitive Toolkit	1.6%	3.0%	-45.5%
Theano	1.6%	4.9%	-67.4%
Torch	0.9%	1.0%	-6.1%
TFLearn	0.7%	1.1%	-34.7%
Caffe	0.6%	1.5%	-58.3%

https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html

Keras basics

- Build the model (network)
  - contains layers with neurons and connections
- Compile the model (model.compile)
  - define loss function and optimizer
- Train the model (model.fit)
  - provide samples with known labels
- Use the model for predictions (model.predict)
  - predict labels of unknown samples

The simplest example

```
# import keras
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
# build model
model = Sequential()
model.add(Dense(50, activation='sigmoid'))
model.add(Dense(1, activation='sigmoid'))
# compile model
model.compile(loss='binary_crossentropy',
    optimizer="adam",metrics=['accuracy'])
# train model
model.fit(samples, labels, epochs=100, batch_size=10)
# use model
predicted = model.predict(sample)
```

Decoder

decoder1.ipynb

- Decodes a binary list:
  - [1,0,0,1,0,0,0,1]
- to the number:
  - 145
- Input: 8 values (0 or 1)
- Output: one value (the number)

pawel.kasprowski@polsl.pl

Deep Learning with Python

Next lecture:
Neural Networks in Keras/Tensorflow

Introduction to Data Mining

Paweł Kasprowski, PhD, DSc.