ICR The Institute of Cancer Research

in partnership with

CANCER RESEARCH UK

The ROYAL MARSDEN NHS Foundation Trust

Support Vector Machine k-fold Cross Validation & Boosting

ICR Sutton Campus

09/05/2018

Priya Narayanan, Shan Raza, Sebastian Pölsterl, Igor Kozin

Materials

https://www.csie.ntu.edu.tw/~cjlin/liblinear/
http://research.microsoft.com/~cmbishop/PRML
Video Lectures of Andrew Ng
https://machinelearningmastery.com/blog/
Data Science Tutorials - All in One (you tube)

Plan

• 25/04: Introduction

02/05: Linear algebra, calculus, least • squares and logistic regression

• 09/05: SVM; k-fold cross-validation and boosting

• 16/05: CNNs; Backprop; Representation Learning; Regularisation; SGD

• 23/05: Image classification using Deep Learning models; Keras, Tensorflow and TF-tensorboard

Please feel free to drop us an e-mail if you have any questions or suggestions.

Recap

- **Prepare Data**: Recipes for data preparation including data cleaning, feature selection and data transforms.
- **Algorithms**: Recipes for using a large number of machine learning algorithms, including linear, nonlinear, and decision trees for classification and regression, SVM for non linear problems.
- **Evaluate Algorithms**: Using re-sampling methods, algorithm evaluation metrics
- and model selection.
- **Improve Results**: Fine tuning, Hyper parameter optimization and ensemble methods.
- **Finalize Model**: Recipes to save and test models.

conda create –n myenv python=3.5 anaconda

Slides & Homework

You can find the presentation and exercises at:

https://github.com/ink1/dl-training

Look for To do for participants in the exercises jupyter files and upload your notebook to solution or email us your solution.

Support Vector Machine

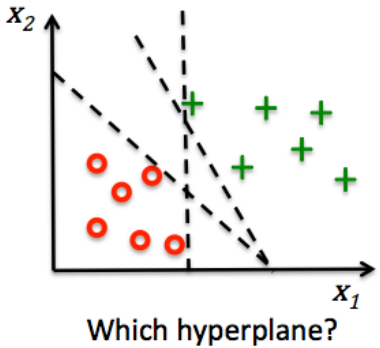
Definition

- SVM are supervised machine learning models with associated learning algorithms that analyse the data used for classification and regression analysis.

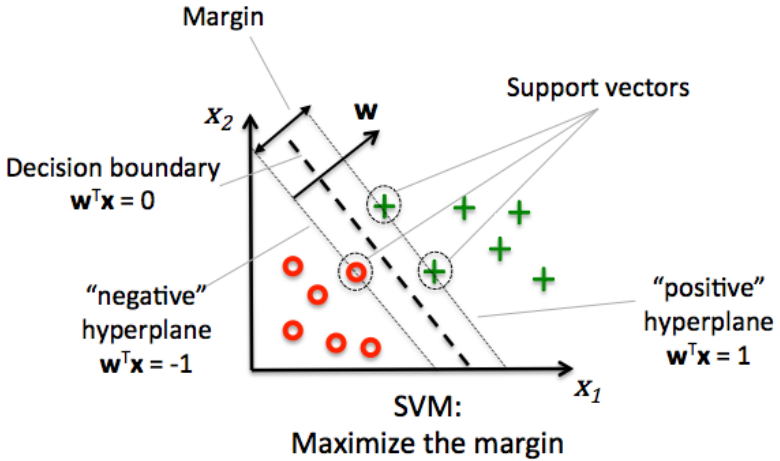
- SVM constructs a hyperplane or set of hyperplanes in a high or infinite-dimensional space and it can be used for classification and regression

- Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class.

Support Vector Machines



Which hyperplane?



SVM:
Maximize the margin

What if we want to split data in best possible way?

Support Vector Machines

Support vector machines can be used both for classification and regression and thanks to the Kernel trick in a wide range of applications. The best choice of Kernel and its parameters is not obvious and requires exhaustive testing for the dataset.

Objective function =

$$min_w \lambda \parallel w \parallel^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+$$

$$\frac{\delta}{\delta w_k} \lambda \parallel w \parallel^2 = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k} \left(1 - y_i \langle x_i, w \rangle\right)_+ = \begin{cases} 0, & \text{if } y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

The first term is a regularizer, the heart of the SVM, the second term the loss. The regularizer balances between margin maximization and loss. We want to find the decision surface that is maximally far away from any data points.

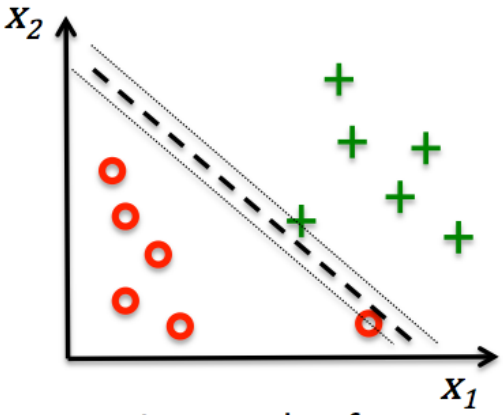
Update rule for Mis-classified sample

$$w = w + \eta(y_i x_i - 2\lambda w)$$

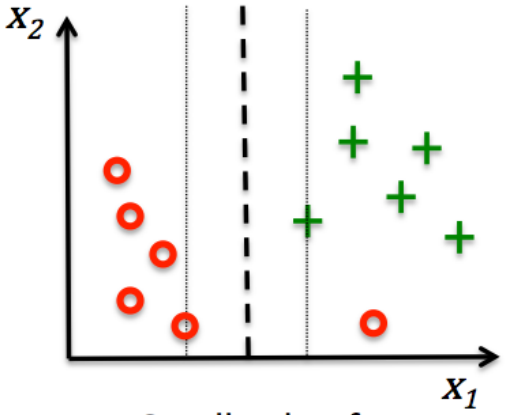
Update rule for correctly classified sample

$$w = w + \eta(-2\lambda w)$$

Support Vector Machines



Large value for parameter *C*



Small value for parameter *C*

Tuning parameters

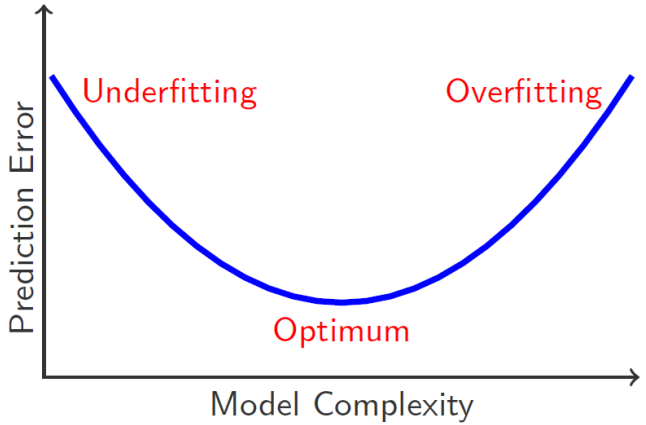
C – regularization parameter

Kernel – ‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’

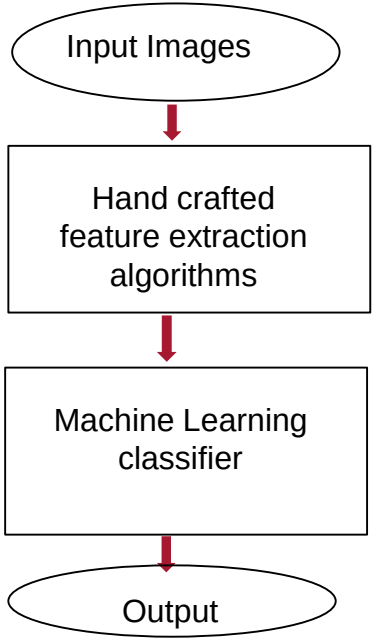
Gamma – kernel coefficient for ‘rbf’, ‘poly’ and ‘sigmoid’

Validation – Over- and Underfitting

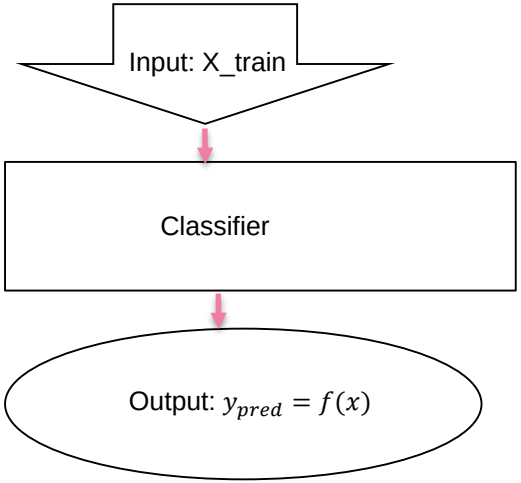
- **Overfitting**: A model with zero or very low training error is likely to perform well on the training data but generalize badly (model too complex).
- **Underfitting**: Model does not capture the underlying structure and hence performs poorly (model too simple).



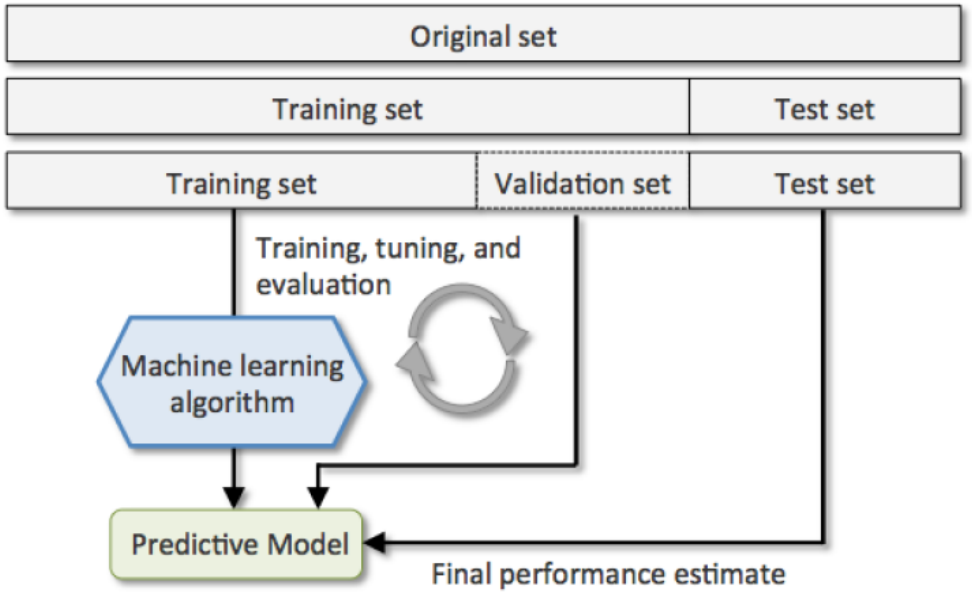
Generalized Pipeline for building applications



Notations
X_train -> Training feature
Y_train -> Training label
X_test -> Test feature
Y_test -> Test label



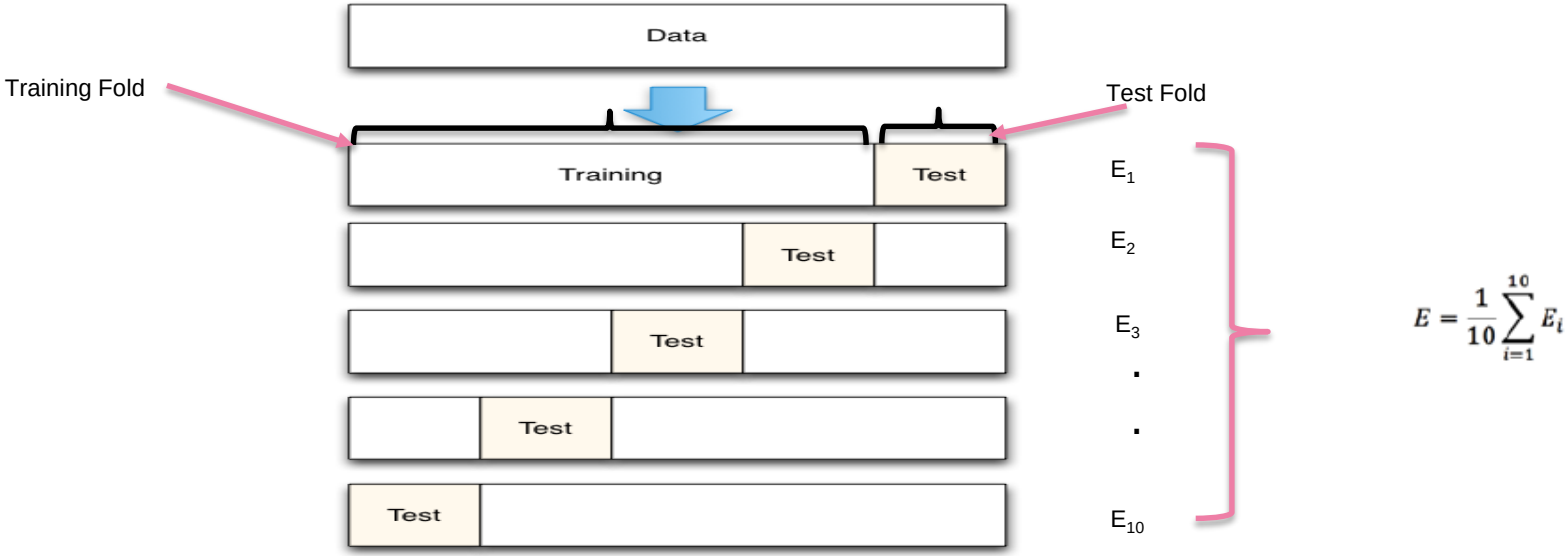
Generalized Pipeline for building applications



Cross Validation

Cross-validation: Split data set into *k* equally large parts.

Stratified cross-validation: Ensures that the ratio between classes is the same in each fold as in the complete dataset.



$$E = \frac{1}{10}\sum_{i=1}^{10} E_i$$

Advantages and Disadvantages of SVM

Advantages

Effective in higher dimensional space

Different kernel functions for various decision functions.

Disadvantages

Poor performance =>Number of features >Number of samples

Exercises

02_svm_classify_compare.ipynb

Experiment validation ideal set up

1. Usually construct a training set, validation set and test set
2. Estimate the prediction error to chose best model (e.g params such as C, gamma for SVMs)
3. Test set is used to assess how well the model generalizes.

In the jupyter files, I have only used train and test split for now.

Evaluation Metrics

Classification

- Confusion Matrix
- Accuracy
- Error rate
- Receiver operating characteristics

Validation

- Leave one out cross validation

Evaluation Metrics

For a binary classification problem:

- **True Positive (TP)** is defined as positive sample **correctly classified** as belonging to the positive class
- **False Positive (FP)** is defined as negative sample **misclassified** as belonging to the positive class
- **True Negative (TN)** is defined as negative sample **correctly classified** as belonging to the negative class
- **False Negative (FN)** is defined as positive sample **misclassified** as belonging to the negative class

Ground truth labels

Predicted Labels	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Evaluation Metrics

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

Error rate = 1 - Accuracy

The true positive rate (TPR) and false positive rate (FPR) are performance metrics that are especially useful for imbalanced class problems:

$$FPR = FP/N = \frac{FP}{FP+TN}$$

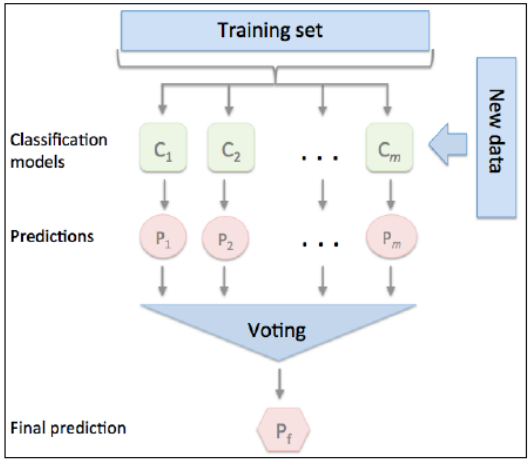
$$TPR = TP/P = \frac{TP}{FN+TP}$$

Overall Summary

- Different performance measures for classification exist.
- ROC and Precision-Recall curves can be applied for binary classifiers and it return probabilities or scores.
- Cross-Validation is used in validation to estimate performance of the model and derive the optimal model.

Majority Voting Classifier

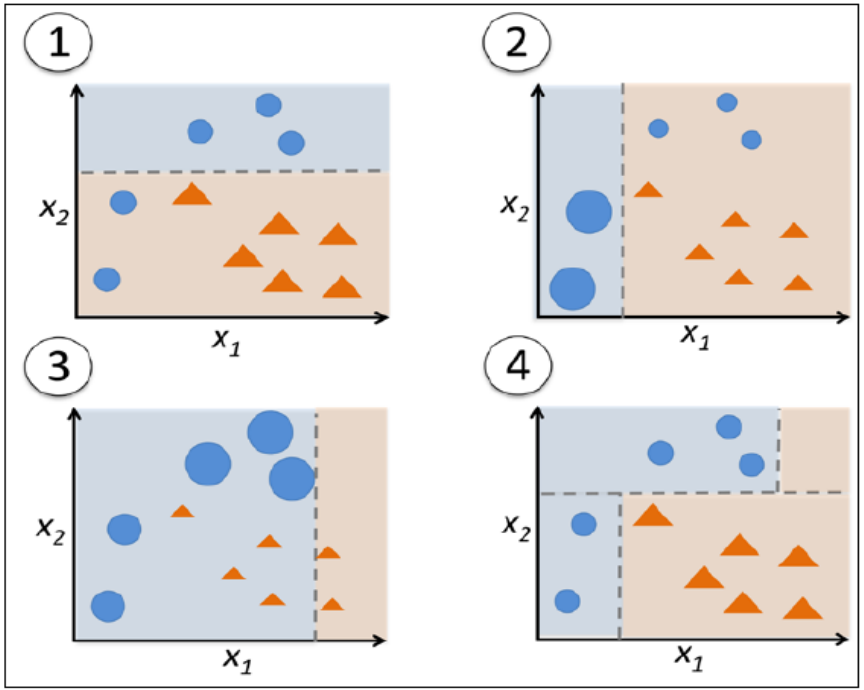
$$\hat{y} = \arg\max_i \sum_{j=1}^{m} w_j \chi_A \left(C_j(\boldsymbol{x}) = i \right)$$



Boosting

- AdaBoost Classifier

An AdaBoost classifier is a meta-estimator This starts by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

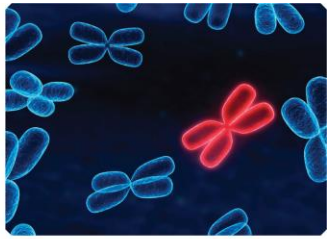


Note: Sampling training dataset without replacement


ICR

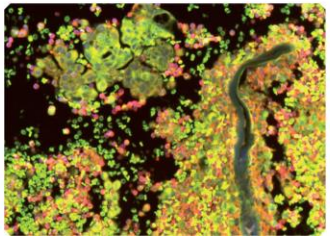
Unrivalled track record

ICR The Institute of Cancer Research



Making the discoveries that defeat cancer

ICR



One of the world's most influential cancer research institutes