Predictive Modelling

Classification - K Nearest Neighbours

Jonathan Mwaura

Khoury College of Computer Sciences

July 23, 2024

The Roux Institute
at Northeastern University

Introduction

Textbook

Reading: Chapter 4 of: Gareth James et al (2021) . An Introduction to Statistical Learning (2nd Edition) .
`https://www.statlearning.com/`

Acknowledgements

These slides have been adapted from the following Professors:

1) Andrew Ng - Stanford
2) Eric Eaton - UPenn
3) David Sontag - MIT
4) Alina Oprea - Northeastern

Supervised Learning

Problem Setting

- Set of possible instances $\mathcal{X}$
- Set of possible labels $\mathcal{Y}$
- Unknown target function $f : \mathcal{X} \rightarrow \mathcal{Y}$
- Set of function hypotheses $H = \{h \mid h : \mathcal{X} \rightarrow \mathcal{Y}\}$

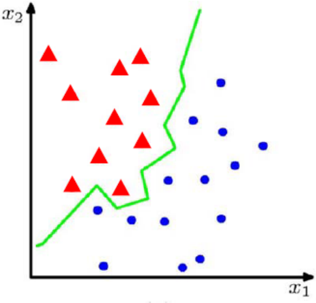
Input: Training examples of unknown target function f

$$\{x_i, y_i\}, \text{for } i = 1, \dots, N$$

Output: Hypothesis $\hat{f} \in H$ that best approximates f

$$\hat{f}(x_i) \approx y_i$$

Classification



- Suppose we are given a training set of N observations

$$\{x_1, \dots, x_N\} \text{ and } \{y_1, \dots, y_N\}, x_i \in R^d, y_i \in \{0, 1\}$$

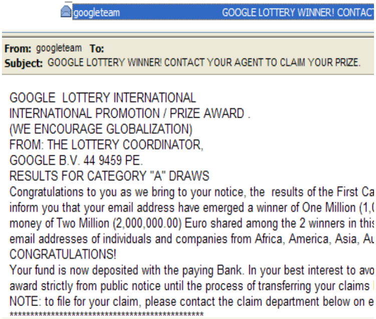
Binary or discrete

- Classification problem is to estimate f(x) from this data such that

$$f(x_i) = y_i$$

Example 1: Binary classification

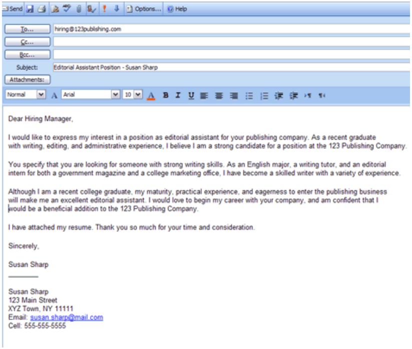
Classifying spam email



googleteam GOOGLE LOTTERY WINNER! CONTAC

From: googleteam **To:**
Subject: GOOGLE LOTTERY WINNER! CONTACT YOUR AGENT TO CLAIM YOUR PRIZE.

GOOGLE LOTTERY INTERNATIONAL
INTERNATIONAL PROMOTION / PRIZE AWARD .
(WE ENCOURAGE GLOBALIZATION)
FROM: THE LOTTERY COORDINATOR,
GOOGLE B.V. 44 9459 PE.
RESULTS FOR CATEGORY "A" DRAWS
Congratulations to you as we bring to your notice, the results of the First Ca
inform you that your email address have emerged a winner of One Million (1,(
money of Two Million (2,000,000.00) Euro shared among the 2 winners in this
email addresses of individuals and companies from Africa, America, Asia, Au
CONGRATULATIONS!
Your fund is now deposited with the paying Bank. In your best interest to avo
award strictly from public notice until the process of transferring your claims
NOTE: to file for your claim, please contact the claim department below on e
****



To... hiring@123publishing.com

Subject: Editorial Assistant Position - Susan Sharp

Normal Arial 10

Dear Hiring Manager,

I would like to express my interest in a position as editorial assistant for your publishing company. As a recent graduate with writing, editing, and administrative experience, I believe I am a strong candidate for a position at the 123 Publishing Company.

You specify that you are looking for someone with strong writing skills. As an English major, a writing tutor, and an editorial intern for both a government magazine and a college marketing office, I have become a skilled writer with a variety of experience.

Although I am a recent college graduate, my maturity, practical experience, and eagerness to enter the publishing business will make me an excellent editorial assistant. I would love to begin my career with your company, and am confident that I would be a beneficial addition to the 123 Publishing Company.

I have attached my resume. Thank you so much for your time and consideration.

Sincerely,

Susan Sharp

Susan Sharp
123 Main Street
XYZ Town, NY 11111
Email: susan.sharp@mail.com
Cell: 555-555-5555

Content-related features

- Use of certain words
- Word frequencies
- Language
- Sentence

Structural features

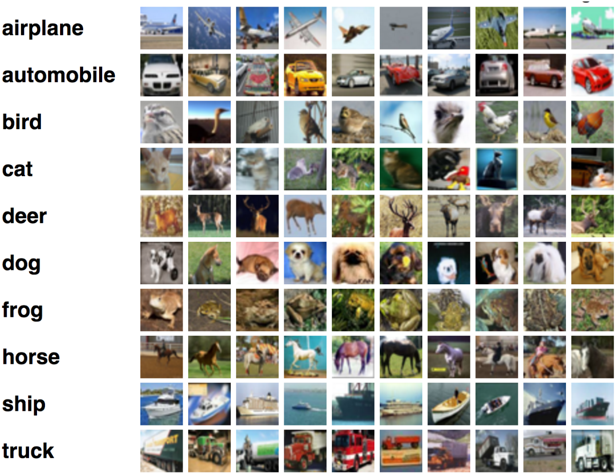
- Sender IP address
- IP blacklist
- DNS information
- Email server
- URL links (non-matching)

The Roux Institute
at Northeastern University

Binary classification: SPAM or HAM

Example 2: Multi-class classification

Image classification



Multi-class classification

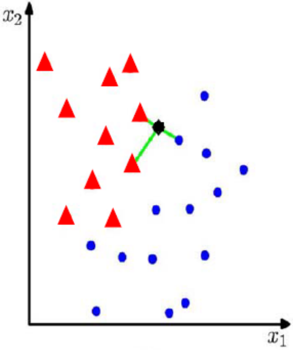
K Nearest Neighbour (K-NN) Classifier

Algorithm

- For each test point, x, to be classified, find the K nearest samples in the training data
- Classify the point, x, according to the majority vote of their class labels

e.g. K = 3

- applicable to multi-class case



The Roux Institute
at Northeastern University

Distance Metrics

- Euclidean Distance

$$\sqrt{\left(\sum_{i=1}^{k}(x_i - y_i)^2\right)}$$

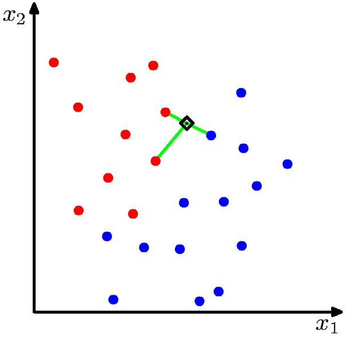
- Manhattan Distance

$$\sum_{i=1}^{k}|x_i - y_i|$$

- Minkowski Distance

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{\frac{1}{q}}$$

kNN

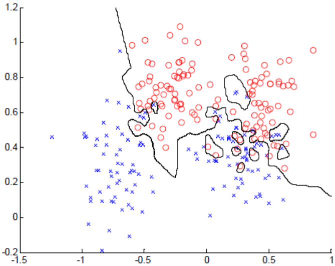


- Algorithm (to classify point x)
  - Find k nearest points to x (according to distance metric)
  - Perform majority voting to predict class of x
- Properties
  - Does not learn any model in training!
  - Instance learner (needs all data at testing time)

K = 1

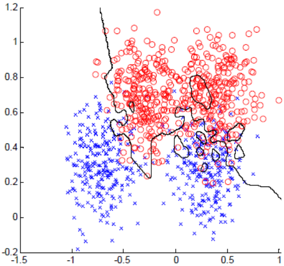
Overfitting!

Training data



error = 0.0

Testing data

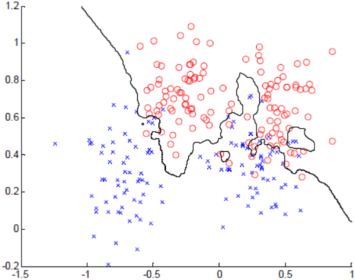


error = 0.15

How to choose k (hyper-parameter)?

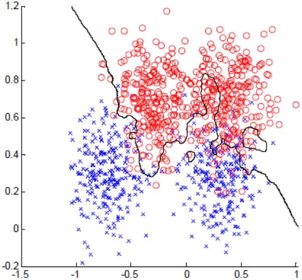
K = 3

Training data



error = 0.0760

Testing data

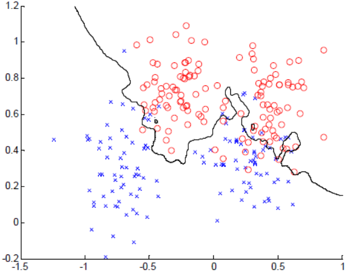


error = 0.1340

How to choose k (hyper-parameter)?

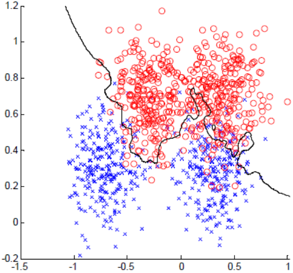
K = 7

Training data



error = 0.1320

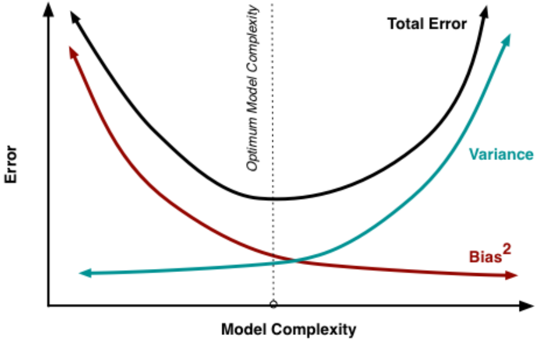
Testing data



error = 0.1110

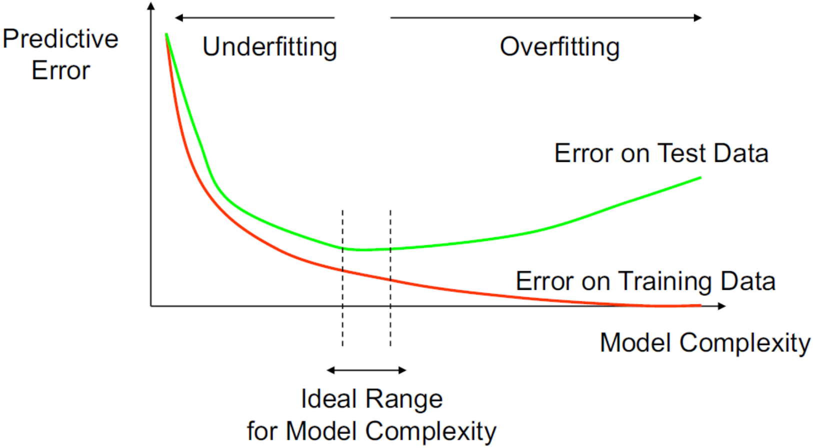
How to choose k (hyper-parameter)?

Bias-Variance Tradeoff for kNN



K decreases

How Overfitting Affects Prediction



How can we avoid over-fitting without having access to testing data?

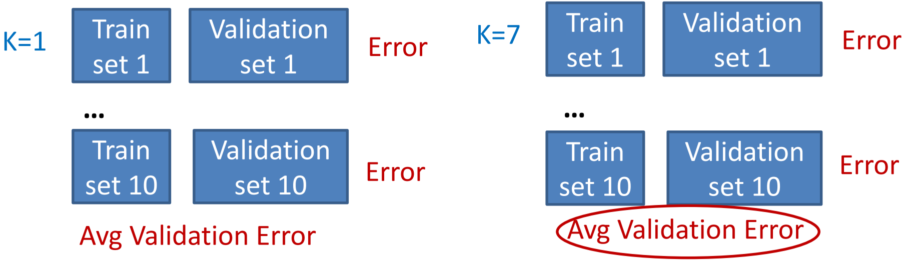
Cross Validation

As K increases:

- Classification boundary becomes smoother
- Training error can increase

Choose (learn) K by cross-validation

- Split training data into training and validation
- Hold out validation data and measure error on this



K=1

Train set 1 | Validation set 1 | Error

...

Train set 10 | Validation set 10 | Error

Avg Validation Error

K=7

Train set 1 | Validation set 1 | Error

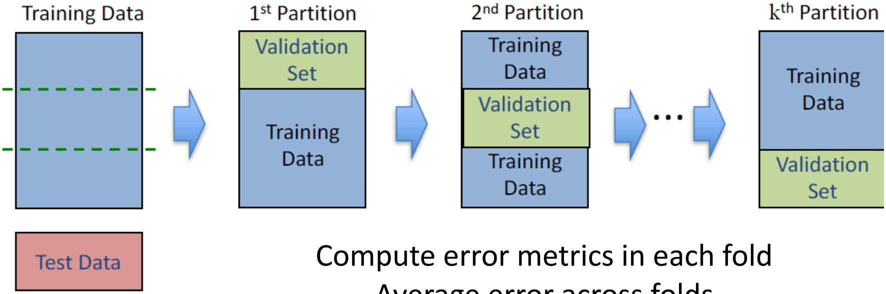
...

Train set 10 | Validation set 10 | Error

Avg Validation Error

Minimum error!

Cross Validation

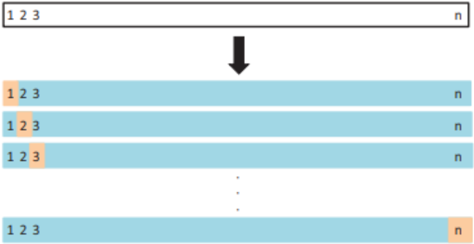


Compute error metrics in each fold
Average error across folds

1. k-fold CV

- Split training data into k partitions (folds) of equal size
- Pick the optimal value of hyper-parameter according to error metric averaged over all folds

Cross Validation



2. Leave-one-out CV (LOOCV)

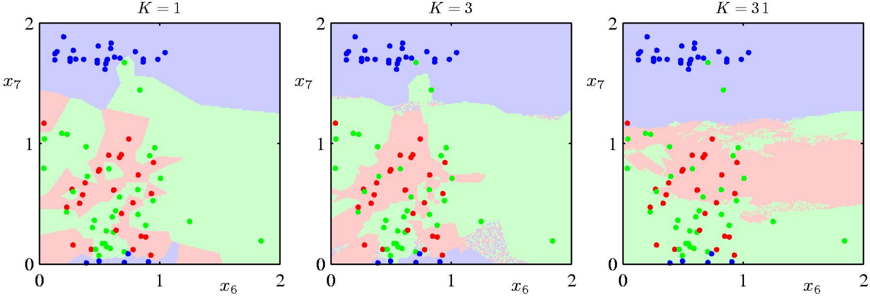
   – k=n (validation set only one point)

- Pros: Less bias
- Cons: More expensive to implement, higher variance
- Recommendation: perform k-fold CV with k=5 or k=10

Cross-Validation Takeaways

- General method to estimate performance of ML model at testing and select hyper-parameters
  - Improves model generalization
  - Avoids overfitting to training data
- Techniques for CV: k-fold CV and LOOCV
- Compare to regularization
  - Regularization works when training with GD
  - Cross-validation can be used for hyper-parameter selection
  - The two methods can be combined

K-Nearest-Neighbours for Multi-class Classification



Vote among multiple classes