Karina Palyutina

Machine learning inference of search engine heuristics

Part II Project

St Catharine's College

March 22, 2013

Proforma

Name:	**Karina Palyutina**
College:	**St Catharine's College**
Project Title:	**Machine learning inference of search engine heuristics**
Examination:	**Part II Project**
Word Count:	[1]
Project Originator:	Dr Jon Crowcroft
Supervisor:	Dr Jon Crowcroft

[1] This word count was computed by `detex diss.tex | tr -cd '0-9A-Za-z \n' | wc -w`

Original Aims of the Project

Work Completed

Special Difficulties

Declaration of Originality

I, Karina Palyutina of St Catharine's College, being a candidate for Part II of the Computer Science Tripos , hereby declare that this dissertation and the work described in it are my own work, unaided except as may be specified below, and that the dissertation does not contain material that has already been used to any substantial extent for a comparable purpose.

Signed

Date

Contents

Introduction

This project is inspired by increasing importance of search engine rankings. Today major search engines given a query return web pages in an order determined by secret algorithms. Such algorithms are believed to incorporate multiple unknown factors. For instance, Google claims to have over 200 unique factors that influence a position of a webpage in the search results relative to a query [2]. Only a handful of these factors are disclosed to the webmasters in the form of very general guidelines. Moreover, the Google algorithm in particular is updated frequently. However, most of the knowledge around the area amounts to speculation. Despite the fact that it is possible to pass a vast number of queries through the black box of any existing search engine,the immensity of the search space, and instability of such algorithms make them impossible to reverse engineer.

Machine learning is a natural approach to inferring the true algorithm from a subset of all possible observations. However, applying machine learning techniques to real search engines would be hardly effective, as the dynamic nature of the algorithms and the web as well as lack of meaningful feedback would prevent incremental improvement: when there are as many as 200 features in question, false assumptions made by a learner may have an unpredictable effect on its performance.

More generally, there are certain ambiguities associated with machine learning, which are 'problem-specific'. For example, it proves difficult to decide how much training data is necessary, as well as and selecting it to avoid over/underfitting[1]. Similarly, it is not straightforward which machine learning technique is best for a particular problem.

This project is concerned with application of machine learning techniques to search engines. The aim of the project, in particular, is to explore how machine learning techniques can be used effectively to infer algorithms from search engines. To address the limitations imposed by existing search engines, part of the task is to develop a toy search engine that allows me to comtrol the nature

[2] http://www.google.com/competition/howgooglesearchworks.html

and complexity of used heuristics. Such transparency addresses the problems stated above and, more importantly, allows for useful evaluation of machine learning techniques by providing meaningful feedback.

Even though this study does not attempt to reverse engineer any existing heuristics, the results can be applied to such an ambitious task. Moreover, such a framework is potentially more general and can be used for a range of problems.

TODO: overview of the chapters here.

Chapter 1

Preparation

This chapter describes work that has been done before coding was started. In particular, it focuses on the reasoning behind the design of the system to be implemented. The first section is devoted to research undertaken to determine what can be done and how best to do it. The second section formulates the system requirements, namely formalizes everything that is developed in this project. The last section outlines the particulars of the software engineering approach to be adopted by this project.

1.1 Formulating the Goals

A particular difficulty in this project has been in planning what has to be done. Due to the exploratory nature of the project the course of action had to be predominantly determined by the outcome of a current tactic. Moreover, the unknowns originating from the machine learning further complicated matters.

1.1.1 System Overview

To achieve the goal of the project, a machine learning techniques comparison framework was necessary. In the Introduction I mentioned the benefit of having a transparent system as an object of learning. To further justify this decision, it is worth mentioning that generalisation using machine learning is different from most optimization problems in that the function that is being optimized is out of our reach, and all that is visible to the machine learner is the training error. Because our goal is not the correct classification of real data, but identifying the means to correct classification, it is important that informed choices are made towards improvement of the learner. Taking this into account, knowing

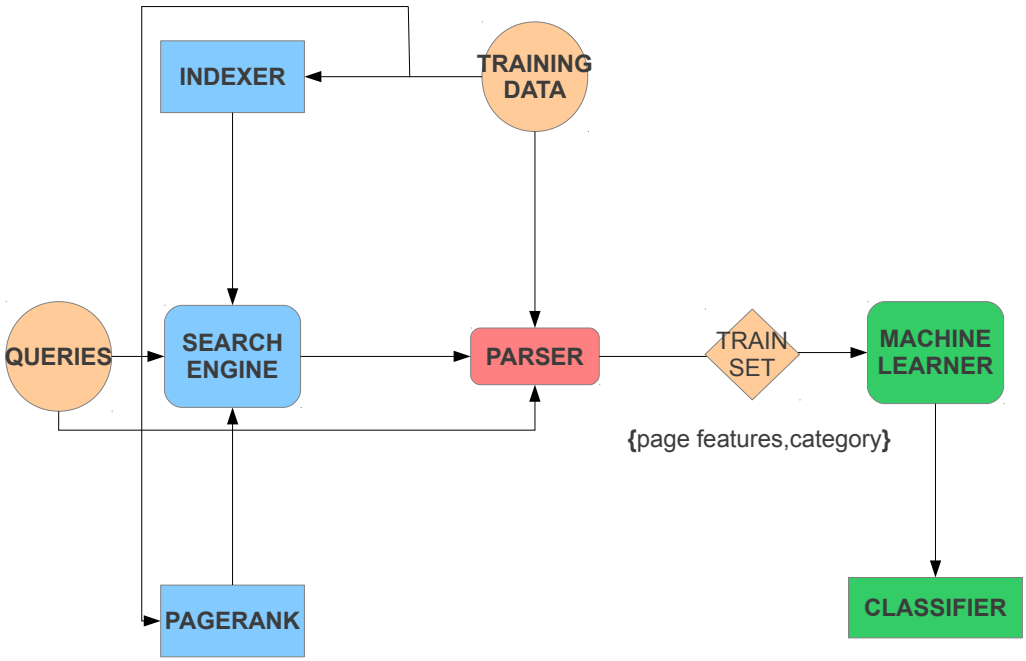


Figure 1.1: Overview of the system. Three major parts from left to right are search engine, parser and machine learner.

the function that we want to learn and having direct control over it will guide the improvement of the search engine.

This argument motivates a system in three parts: a search engine, a machine learner and a parser to mediate between the two. Figure 1.1 illustrates the proposed learning system. Training data is a set of web pages set aside specifically for training purposes.

Choice of Programming Language

When choosing a programming language, main considerations reduced to library availability and simplicity. The project imposes no special requirements on the language, apart from, perhaps, library infrastructure for parsing web pages. Python is simple language with extensive library support. As for efficiency, all the mathematical operations in this project rely on python math libraries, which are implemented in C. I have not programmed in Python before the project, so a slight overhead was caused by having to learn a new language.

Data

Web pages used as Training and Test data are not required to have any special properties, but diversity and typicality are seen as advantageous. As for the size of the training data, Domingos [1] suggests that a primitive learner given more data performs better than a more complex one with less data. This, of course, is under certain assumptions of data quality, namely the assumption that the training data is a representative subset of all the possible data. Intuitively, provided there is no bias in data gathering, more data implies better generality. I have started with a training set spanning an order of a few thousands of pages, however, in practice, I found that there is no particular improvement beyond a thousand pages. **TODO: Link to relevant part or example data here?**

1.1.2 Search Engine

Next important decision regarded the search engine. Originally, I considered using open source existing engines, in particular, Lucene. Even though I could freely modify it for the purposes of the project, the complexity of it was superfluous. I saw writing a simple search engine as a more beneficial exercise, as developing it in the first place potentially gives an insight into the problem.

Functionally our search engine is a black box that takes a set of webpages and a set of queries and outputs an order. The order is determined by the features of the page, which together make up a score. The score is the function we want to infer using the ranking assigned by the search engine, however, we are only given the order as evidence. Machine learning paragraph below will address this issue in more details.

In general there are two aspects of information retrieval that have to be accounted for: precision and recall. Precision is the fraction of retrieved pages that are relevant to the query, whereas recall is the fraction of relevant documents that are retrieved. Even though both are important for a good search engine, but in practice, the web is very large, and so precision, or even *precision at* n[1]) has become more prominent in defining a good search engine: very rarely the user actually browses returns that are not in the top few tens of returned pages. Therefore, modern search engines tend to focus on high precision at the expense of recall [5]. Therefore, we will concentrate primarily on precision, when designing a search engine.

[1] 'Precision at n' only evaluates precision with respect to n topmost returned pages.

PageRank

The PageRank algorithm was originally described in a research paper by Page and Brin[6]. It was first introduced as a way 'to measure the relative importance of web pages'. PageRank is interesting to include in this project not only because it is a defining feature of the Google search engine, but because it is a unique feature of its type, as it depends on the link structure of the whole web as opposed to other web features such as word count and alike.

The basic idea is to capture the link structure of the web and provide a ranking of pages that is robust against manipulation. To achieve this, the importance 'flows' forward: the backlinks share their importance with the pages they link to. Such a simplified ranking of a set of pages can be expressed as an assignment

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v} \tag{1.1}$$

An intuitive justification for such ranking is by analogy with a citation network, we are likely to find highly cited pages more important. The equation is recursive, so iterating until convergence results in a steady state distribution, which corresponds to the PageRank vector. Together with this intuition the paper introduces the *Random Surfer Model*: we are interested in a steady state distribution of a random walk on the Web graph. At each step the surfer either follows a random link or 'teleports' to a random page. The paper justifies the approach of ignoring the dangling links, as it doesn't have a significant affect on the ranking.

The teleportation vector determines whether the PageRank is personalized. For the non-personalized version the teleportation vector holds equal probabilities for all pages in the Web, whereas in a personalized approach the probabilities are distributed according the knowledge of the surfer's previous activity. In this project we are only concerned with nonpersonalized ranking, which simplifies things a little.

1.1.3 Machine Learning

I have now covered the main peripheral decisions, but it is machine learning that constitutes the central part of the project. The field was completely new to me to start with, so research of different techniques was a big part of the preparation.

Machine learning is a vast field and very little is prescribed. The supervised learning approach, where the hidden function is inferred from the labelled training data, best fits our purposes for a number of reasons. Firstly, it applies well

to the real-world problem of inferring the heuristics of the real search engines, as we can select training data and obtain the labels simply by querying. Secondly, supervised learning is widely used and offers a variety of flexible techniques.

An interesting known problem with machine learning in general is referred to as the Curse of Dimensionality. In the real world scenario – suppose if we were inferring an existing search engine from the observed query responses – we could not be sure about which features are and which are not relevant to the hidden algorithm. Potentially, some features we might choose to use will increase the dimensionality, but would not actually be used by the real algorithm. Gathering more features can hurt, as it makes the learner infer nonexistent dependencies. Clearly within the scope of this project we will know at each point which features are or are not used. So observations can be made as to how using more features than the search engine will affect the classifier.

Another major issue that is common to all machine learning methods is over/under-fitting. These refer to the problem of finding balance between the generalized model and the training data at hand. This is yet another source of interest to this project, as we can observe the behaviour of the learner when certain control parameters are changed.

It is generally recommended that the simplest learners are tried first[1]. Of all learners Naive Bayesian is one of the most comprehensible.This in itself is a major advantage according to the Occam's razor principle, which finds ample application in machine learning. Hence, we start with describing the principles of the two machine learning techniques used - Naive Bayes and Support Vector Machines.

Naive Bayes

Naive Bayes is a probabilistic classifier based on the Bayes Theorem. The posterior probability $P(C|\vec{F})$ denotes the probability that a sample page with a feature vector $\vec{F} = (F_1, F_2, \ldots, F_n)$ belongs to class C. The posterior probability is computed from the observable in the training data: the prior probability $P(C)$ – the unconditional probability of a page belonging to the class C, the likelihood $P(\vec{F}|C)$ and the evidence $P(\vec{F})$:

$$P(C|\vec{F}) = \frac{P(C)P(\vec{F}|C)}{P(\vec{F})} \tag{1.2}$$

The simplicity of Bayesian approach owes to the conditional independence assumption: each F_i in $\vec{F}$ is assumed to be independent of one another to get $P(\vec{F}|C) = P(F_1|C) * P(F_2|C) * \cdots * P(F_n|C)$. This leads to a concise classifier

definition:

$$\hat{C} = argmax_C P(C) \prod_{i=1}^{n} P(F_i|C) \tag{1.3}$$

where C is the result of classification of a page with feature vector $F_1, F_2, \ldots, F_n$.

In practice, the crude assumption rarely holds and is likely to be violated by our data, as we expect features of pages to be interdependent. However, it has been shown that Naive Bayes performs well under zero-one loss function in presence of dependencies[2]. This has a few implications for this project, particularly, on evaluation methods

As we have seen, Naive Bayes assigns probabilities to possible classifications in the process of classifying. Even though it generally performs well in classification tasks, these probability estimates are poor [3]. However, despite poor probability estimates, there exist several frameworks, which make use of Bayesian classification and achieve decent performance in ranking. For example, Zhang [8] experimentally found that Naive Bayes is locally optimal in ranking. The paper defines a classifier as locally optimal in ranking a positive example E if there is no negative example ranked after E and vice versa for a negative exaple. A classifier is global in ranking if it is locally optimal for all examples in the example set: in other words, it is optimal in pairwise ranking. It is particularly interesting that the paper discovered that Naive Bayes is globally optimal in ranking on linear functions that have been shown as not learnable by Naive Bayes[2]. Another framework for ranking [7] is based on Placket-Luce model, which reconciles the concepts of score and rank. This framework is based on minimizing the Bayes risk over possible permutations.Existence of such frameworks suggest that Naive Bayes is an adequate choice for a prototype learner.

Although classification is a useful technique to try, it feels more natural to represent our score or rank as real numbers rather than classes. Regression is another approach to machine learning and the next learning technique we explore – Support Vector Regression – is non-probabilistic, to try something as distant from Naive Bayes as possible.

ϵ-Support Vector Regression

While the binary classification problem has as its goal the maximization of the margin between the classes, regression is concerned with fitting a hyperplane through the given training sequence.

[2] *M-of-N Concepts* and *Conjunctive Concepts* can't be learnt by Naive Bayes classifier but can be optimally ranked by it according to Zhang [8].

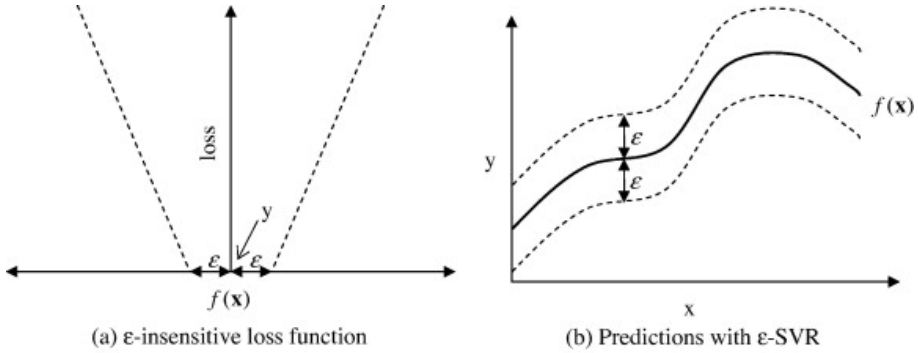


Figure 1.2: TODO: plot these yourself!

A great advantage of Support Vector machines is that they can perform non-linear classification or regression by using what is referred to as the *Kernel Trick* - an implicit mapping of features into a higher dimensional space, in which the data is linearly separable. The choice of a kernel function is problem specific and the best one is usually decided by experiment.

We begin with the theoretical foundations of Support Vector Regression, which were first proposed by Vapnik **??**.
Define a training sequence as a set of training points $D = \{(\mathbf{x_1}, t_1), (\mathbf{x_2}, t_2), ..., (\mathbf{x_l}, t_l)\}$ where $\mathbf{x_i} \in R^n$ is a feature vector holding features of pages and $\mathbf{t_i} \in R$ is the corresponding ranking of each page.

In simple linear regression the aim is to minimize a regularized error function. We will be using an ϵ-insensitive error function(see Figure 1.2 (a)).

$$E_\phi(y((x) - t)) = \begin{cases} 0 & \text{if } |y(\mathbf{x}) - t)| < \epsilon \\ |y(\mathbf{x}) - t)| - \epsilon & \text{otherwise} \end{cases}$$

where $y((x) = \mathbf{w^T}\phi(\mathbf{x}) + b$ is the hyperplane equation (and so $y(\mathbf{x})$ is the predicted output) and t_n is the target (true) output.

The regression tube then contains all the points for which $y(\mathbf{x_n}) - \epsilon \leq t_n \leq y(\mathbf{x_n}) + \epsilon$ as shown in Figure 1.2(b).

To allow variables to lie outside of the tube, slack variables $\xi_n \geq 0$and $\xi_n^* \geq 0$ are introduced. The standard formulation of the error function for support vector regression (ref Vapnik 1998) can be written as follows:

$$E = C \sum_{n=1}^{N} (\xi_n + \xi_n^*) + \frac{1}{2}\|\mathbf{w}\|^2 \tag{1.4}$$

E must be minimized subject to four constraints:

$$\xi_n \geq 0, \tag{1.5}$$
$$\xi_n^* \geq 0, \tag{1.6}$$
$$t_n \leq y(\mathbf{x_n}) + \epsilon + \xi_n, \tag{1.7}$$
$$t_n \geq y(\mathbf{x_n}) - \epsilon - \xi_n^*, \tag{1.8}$$

This constraint problem can be transformed into its dual form by introducing Lagrange multipliers $a_n \geq 0, a_n^* \geq 0$. The dual problem involves maximizing

$$L(\mathbf{a}, \mathbf{a}^*) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - a_n^*)(a_m - a_m^*)K(\mathbf{x_n}, \mathbf{x_m}) \tag{1.9}$$

$$-\epsilon\sum_{n=1}^{N}(a_n + a_n^*) + \sum_{n=1}^{N}t_n(a_n - a_n^*)$$

where $K(x_n, x_m)$ is the kernel function, t_n is the target output,

subject to constraints

$$\sum_{n=1}^{N}(a_n - a_n^*) = 0, \tag{1.10}$$
$$0 \leq a_n,\ a_n^* \leq C,\ \ n = 1, ..., l \tag{1.11}$$

Evaluation methodology

Measuring performance of machine learners accurately is a major challenge primarily because both training and testing are dependent both on the quantity and quality of data used. This fact motivates repeating experiments with a variety of data (**TODO: n fold sampling**). Another difficulty arises from the degrees of freedom that are internal to the learners: the hyperparameters. Especially SVM kernels are vulnerable to mis-evaluation due to the number of parameters which need to be set. The issues of hyperparameter tuning and data sampling will be addressed in the Evaluation chapter later. This section briefly touches upon evaluation models exploited.
Two proposed techniques – Naive Bayes and SVM regression – are quite different, so comparing them is potentially erroneous. However, comparisons can be

done within each method, as there is a lot of scope for variety of implementations in each.
As a baseline for the Bayesian approach, a very primitive ranking model will be used. We will simply disregard the scoring function behind the rank and directly infer the ranking function instead. Precision and Recall metric is used to quantify performance. More sophisticated ranking models can then be compared to this basic performance. In particular, weak ranking property can be evaluated.**TODO: explain** Figure 1.3 shows a simplified evaluation framework for Naive Bayes. The Test Data is the data carefully set aside at the beginning that is never exposed to the learner. **TODO: bayes and heuristics?!**

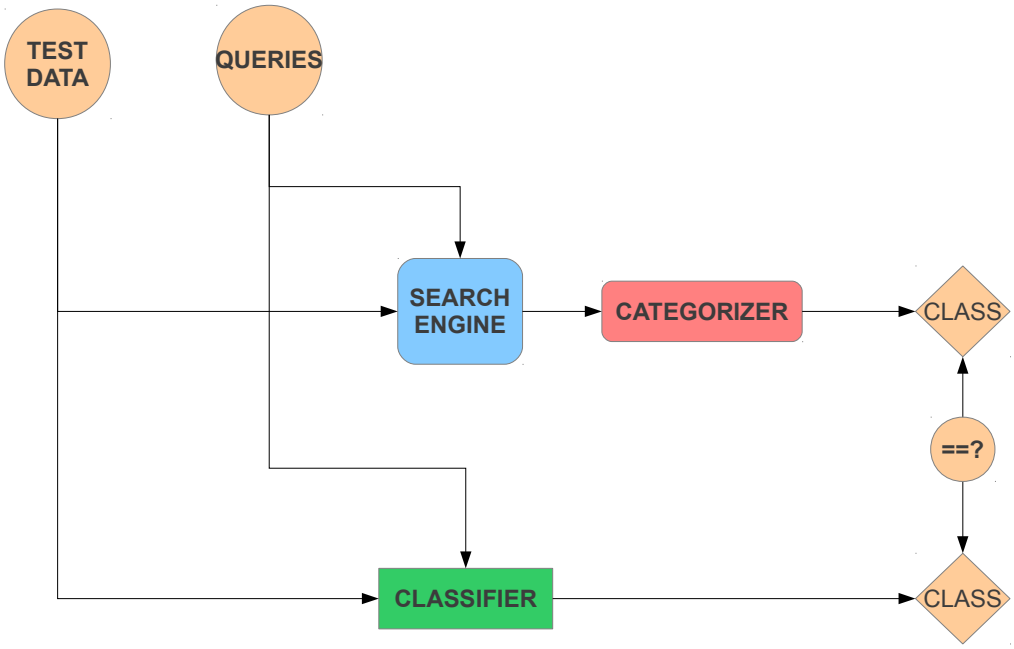


Figure 1.3: Evaluation process.

For an SVM learner the baseline can be set to the performance using a linear kernel below.

$$K(x, y) = x \cdot y$$

where '$\cdot$' denotes the dot product. This is expected to be very high for linear heuristics but a lot lower for non-linear ones. Mean Squared Error will be used

as the risk function to minimize.

$$MSE = \frac{1}{|Actual|}\sqrt{\sum_{i=1}^{N}(Actual_i - Predicted_i)^2}$$

where $Actual$ is an array of true values of size $|Actual|$ and $Predicted$ is the corresponding array of predictions made by the classifier.

1.2 Development Strategy

While the set up of Part II projects encourages waterfall-like development model, this project takes an iterative approach. The first iteration renders a prototype: a primitive search engine with a Naive Bayesian baseline classifier. The next iteration modifies each part of the system towards a more complex solution. Evaluation is performed at each iteration. Within each iteration the development follows the evolved waterfall model – the incremental build model. Each increment represents a functionally separate unit of the system: a search engine, a learner, a parser and an assessment module. Increments are developed sequentially and regression testing is performed separately before integration.

The backup of the code is twofold: every time a substantial change is made a remote version control repository is updated to hold the newest version. Regularly both the code and the data used and obtained during evaluation are also backed up onto an external hard drive.

During the development of the learners a development pool of pages will be used, which must be disjoint with the Training or Test data. This ensures that no optimization is tailored to the data used for evaluation.

Requirements Analysis

The resulting system must comply with the functional and non-functional requirements, but not the non-requirements.

Functional and Non-Functional Requirements:

- A search engine, which given a query must return a relevant subset of pages within 10 seconds in an order corresponding to a given heuristic, which can also be supplied to the search engine.
- Search engine must base ranking decisions on both dynamic (query dependent) and static (query independent) page features.

- A machine learner must be sufficiently isolated from the search engine through an encapsulating interface to avoid the possibility of undesirable interaction.
- Machine learners must be able to process thousands of pages in a reasonable time.
- The evaluation module must write results to persistent storage.

Non-Requirements:

- Usability. This system is not meant to be used by other people and is result oriented, so no user interface is necessary.

Chapter 2

Implementation

This section describes parts of the system that I have implemented. **TODO: overview**

2.1 Data

Data gathering is of major importance, as data decisions have tremendous impact on the rest of the implementation and the success of the project. The rest of the system assumes the existence of a pool of web pages available offline for fast indexing and parsing. Ideally the corpus of web pages would be 'representative' of the whole web, to make generalisations more accurate. Originally I was hoping to get such collection from a resource, like, for example the web corpus of the *Text Retrieval Conference* (TREC). However, such data is not easily available, so I had to retrieve pages using *Wget*. To approach the similarity to the web, I restricted the corpus to one semantic area by only downloading websites related to construction materials[1]. This limitation of topic has certain advantages: relatively few pages need to be downloaded before the pages appear sufficiently interlinked, so the corpus has a distinct structure (as web pages similar in the field link to each other).

To conform with machine learning guidelines I have originally separated data for Test and Training. Different seed pages were used, as well as each resides in their own directory. Each directory is estimated at around 3000 pages. In addition to this, I have taken the further precaution of setting aside a small Development corpus to be used only while development to ensure no bias towards training

[1]There is no particular reason to have chosen this particular topic, except it has been used by the external project originator in his experiment trying to infer Google's ranking factors. This is described in the Proposal in more detail.

data at the implementation stage. Also, its comparatively small size meant faster testing. All the links are converted, so that the link structure is preserved.

2.2 Search Engine

The first basic block of the system – the search engine – can be logically split into two main parts: an indexer and a sorter. Because we concentrate on precision, as discussed in the Preparation chapter, we will assume for simplicity that all relevant documents are returned. This allowed me to use an existing library implementation of the indexer and focus on the sorter. This crude assumption also distinguishes the project's goals from Informational Retrieval: we can ignore the evaluation of the search engine itself and move away from the specifics of web pages towards the more general problem, so that the nature of features makes little difference and the existence of a hidden function is all that matters.

Indexer

The requirements on the indexer include flexibility, speed of indexing and retrieval as well as simplicity and usability. The *Whoosh* python library provides all of these, so I used it to build an indexer. Whoosh is an open source indexing library, so I had the option of modifying any part of it. It is built in pure Python, so it has a clean pythonic API. Its primary advantage is fast indexing and retrieval, although we are mostly concerned with retrieval speed, as indexing is done rarely. The predecessors of Whoosh have served as the basis of well-known browsers such as *Lucene*, so it is also a powerful indexing tool, should I have needed more sophistication.

I have defined a very simple schema for indexing. Perhaps, one notable detail is that Whoosh can store timestamps with the index, which enabled me to provide both clean and incremental index methods. The incremental indexing relies on the time stamp stored with the index and compares it to the last-modified time provided by the file system. The user can specify whether indexing has to be done from scratch or updated to accommodate some document changes or document addition/deletion. I haven't originally expected to need an incremental indexing capability, but throughout the project it has permitted for a significant speedup.

Previously, I have defined the sorter as a logical unit that provides ranking to the retrieved documents. For the first prototype, the sorter returned the pages in the order of decreasing PageRank. Subsequently, the sorting has been

decoupled from retrieval and incorporated into the Parser module, which is described later in detail.

PageRank

The PageRank vector is computed using matrix inversion. All matrix operations were performed with the help of the python numerical library 'numpy'. Take t to be the teleportation probability, s=1-t is the probability of following a random link, E is the teleportation probability: equiprobable transitions, as

using non-personalized PageRank (see Preparation), $E_i, j = 1/N$ for all i, j. G is a stochastic matrix holding the link structure of the data, such that

$$G_i, j = \begin{cases} 1/L & \text{there is a link from i to j and L = number of links from i} \\ 0 & \text{there is no link from i to j} \end{cases}$$

Then M is a stochastic matrix representing the web surfer activity, such that M_i, j is the probability of going from page i to page j,

$$M = s * G + t * E \tag{2.1}$$

In one step the new location of the surfer is described by the distribution Mp. We want to find a stationary distribution p, so must have

$$p = M * p \tag{2.2}$$

Substituting 2.1 into 2.2

$$p = (s * G + t * E) * p = s * G * p + t * E * p \tag{2.3}$$

Rearranging equation 2.3 gives

$$p * (I - s * G) = t * E * p \tag{2.4}$$

where I is the identity matrix

We can express E*p as P where P is a vector $[\overbrace{1/N, 1/N, \ldots, 1/N}^{N}].T$,as members of p must sum to one. So computing PageRank amounts to

$$p = t * (I - s * G)^{-1} * P \tag{2.5}$$

where $(I - s * G)^{-1}$ denotes a matrix inverse operation.

This solution is simple at the expense of being slow. Although computing inverse of a matrix is computationally expensive, we don't need to scale beyond

a few thousands of pages. To avoid recomputation, I used python object serialization module - Pickle to store the PageRank vector for each directory. The resultant performance was actually very reasonable, the time spent computing PageRank was insignificantly small in comparison to the time spent crawling the directory.

The PageRank vector computation happens within the `Pagerank` class. The whole `Pagerank` object is written to memory using the python object serialization module `Pickle`. A `load` class method is defined on the `PageRank` class to retrieve the relevant object for a given directory. The class is instantiated with an instance of the `Crawler` class, which embodies the link structure of a directory and is described in the next section.

Crawler

The `Crawler` abstracts away the underlying data directories and computes their link structures as matrices. The matrix G, used for the PageRank computation, represents random link following activity. To obtain such a link structure each page has to be parsed, and all links recorded. Because our data is obtained from a single source page by recursive link following, every page in a directory is guaranteed to be discovered by a spider.

The Crawler class recursively traverses the pages depth first starting with the seed page, the same as the seed page used for recursively downloading the pages from the web. To make sure each page is only explored once, a dictionary is used to hold pairs of absolute path, which uniquely identifies the page, and a numerical value corresponding to the time stamp when the page has been first discovered.

Although every page has a unique path, the links to other pages are relative. Such links need to be normalized to maintain consistency. A page object is used to encapsulate path complexity: all link paths are converted to absolute paths before addition to the dictionary. All outbound links are stored with the page in a Set data structure, such that no link is added more than once.

To produce the stochastic matrix G, we start with an empty NxN matrix, where N is the total number of pages. We assume that whenever a surfer encounters a dangling page – a page that has no outbound links – a teleportation step occurs. Therefore, every dangling page links to every page in the pool including itself with equal probability 1/N. For non-dangling pages, all links are assumed equiprobable and all pages that are not linked to have probability of 0. So if page A links to pages B and C, but not itself or D, its row in G is described by the Table 2.1.

Page	A
A	0
B	1/2
C	1/2
D	0

Table 2.1: Illustration of non-dangling pages: B and C share A's 'importance' equally.

2.3 Machine Learning

In the course of this project two distinct machine learning algorithms have been used. The design goals for the implementations were as usual: speed and correctness. Another important aspect of the design is that the system and the machine learning modules must be sufficiently decoupled. This is an issue of scalability and code reuse. The implication on the machine learner implementations is that both must communicate with the system via the same interface. The rest of the section describes in detail how the two proposed machine learners were implemented.

2.3.1 Naive Bayes

In Section 1.1 it has been shown that Naive Bayesian is a good learner to implement for the first prototype, so a very quick implementation was preferable to make sure the system can potentially function as intended. Due to its simplicity and popularity, Naive Bayesian is widely available in the libraries. Because the implementation of this module is straightforward, not central to the project, I decided to use one of many existing python implementations.

The *nltk* library implementation was particularly appealing as it offers a very concise interface. A classifier object is initialised by the `train` method on the `NaiveBayesClassifier` class. The format of the training set is defined as a list of tuples of featuresets and labels, e.g.
$[(featureset_1, label_1), \cdots, (featureset_N, label_N)]$.

The `train` method simply computes the prior – the probability distribution over labels P(label) and the likelihood – the conditional probability P($featureset = f|label$) by simply counting and recording the relevant instances. The method outputs a `NaiveBayesClassifier` instance parametrized by the two probability distributions. P(features) is not computed explicitly, instead, a normalizing denominator **??** is used.

$$\sum_{l \in labels} (P(l)P(featureset_1|l) \cdots P(featureset_N|l)) \tag{2.6}$$

The `classify` method on the `NaiveBayesClassifier` object takes exactly one featureset and returns a label which maximizes the posterior probability P(label—featureset) over all labels. Previously unseen features are ignored by the classifier, so that to avoid assigning zero probability to everything.

The only tangible difficulty with the implementation was the framework in which classification is done. This task is, however, achieved by the Parser. To batch classify everything in the test directory, the classifier object is passed to the Parser, where the featuresets for the test directories are computed, classified and recorded for the evaluation stage later. The particulars of this process is described in section **??**.

2.3.2 Support Vector Machine

In the Preparaion chapter we have looked at the theory of support vector machines as found in textbooks. This section builds on the theory described and explains further transformations that were an essential part of the implementation.

To implement a support vector machine one must solve a problem of optimizing a quadratic function ssubject to linear constraints – usually referred to as the *Quadratic Programming* (QP) problem. Therefore, the first implementation task was to convert our existing optimization problem into a generic QP form to make use of the available solvers.

The maximization problem 1.9 can be trivially expressed as a minimization problem (equation 2.7).

$$min_{\boldsymbol{\alpha},\boldsymbol{\alpha^*}} \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha^*})^T P(\boldsymbol{\alpha} - \boldsymbol{\alpha^*}) + \epsilon \sum_{i=1}^{l}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{l} t_i(\alpha_i - \alpha_i^*) \tag{2.7}$$

subject to constraints 2.8 and 2.9 below.

$$\mathbf{e}(\boldsymbol{\alpha} - \boldsymbol{\alpha^*}) = 0 \tag{2.8}$$

$$0 \leq \alpha_i, \alpha_i^* \leq C,\ i = 1, ..., l \tag{2.9}$$

where $\mathbf{e} = [1, ..., 1],\ P_{ij} = K(x_i, x_j),\ t_i$ is the target output, $C > 0$ and $\epsilon > 0$.

At this point in the implementation, for the first time, python did not seem like an ideal choice. *Cvxopt* is one of the few python libraries that implements a QP solver. The specification to the QP function is as follows: `cvxopt.solvers.qp(P,q,G,h,A,b)` solves a pair of primal and dual convex quadratic programs

$$\min \frac{1}{2}x^T Px + q^T x \tag{2.10}$$

subject to

$$Gx \leq h \tag{2.11}$$

$$Ax = b \tag{2.12}$$

Described in the next few pages are the transformations I devised to reconcile the minimization problem 2.7 and the library specification 2.10 and their respective constraints.
We take x to encode both $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha^*}$ simultaneously, treating the upper half of x as $\boldsymbol{\alpha}$ and the lower half as $\boldsymbol{\alpha^*}$:

$$x = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha^*} \end{bmatrix}$$

We will see later how this representation allows for elegant representation of the problem (2.7).

First, we express the first term in 2.10 to hold $(\boldsymbol{\alpha} - \boldsymbol{\alpha^*})^T P(\boldsymbol{\alpha} - \boldsymbol{\alpha^*})$. Take matrix P in equation 2.10 as

$$P = \begin{bmatrix} K & -K \\ -K & K \end{bmatrix}$$

where $K_{ij} = K(x_i, x_j)$ is the kernel.
Observe that now

$$x^T Px = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha^*} \end{bmatrix} \begin{bmatrix} K & -K \\ -K & K \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha^*} \end{bmatrix}$$

is equivalent to the first term of equation (1.9)

$$\sum_{n=1}^{N} \sum_{m=1}^{N} (a_n - a_n^*)(a_m - a_m^*) K(x_n, x_m)$$

Now expressing the remaining two terms of 2.7 by taking

$$q = \begin{bmatrix} \epsilon * \mathbf{e} - t_0 \\ \vdots \\ \epsilon * \mathbf{e} - t_{N-1} \\ \epsilon * \mathbf{e} + t_0 \\ \vdots \\ \epsilon * \mathbf{e} + t_{N-1} \end{bmatrix}$$

to achieve

$$q^T x = \epsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) - \sum_{i=1}^{l} t_i (\alpha_i - \alpha_i^*)$$

To encode constraint 2.9 consider the following pair of G and h.

$$G = \begin{bmatrix} \overbrace{-1 & 0 & 0}^{\text{N}} & \overbrace{0 & 0 & 0}^{\text{N}} \\ 0 & -1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \ddots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \ddots & 0 & 0 \\ 0 & \cdots & \cdots & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \ddots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \ddots & 0 & 0 \\ 0 & \cdots & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \} \text{2N} \\ \\ \\ \\ \\ \\ \} \text{2N} \end{matrix}$$

$$h = [\overbrace{0 \dots 0}^{2N} | \overbrace{C \dots C}^{2N}]^T$$

The final constraint 2.8 is trivial to adapt by simply taking

$$A = [\overbrace{1, \dots, 1}^{N}, \overbrace{-1, \dots, -1}^{N}]$$

and

$$b = 0$$

Having worked out all the matrices, the rest of the implementation dealt simply with coding the matrices up. The Cvxopt library has its own matrix constructor, however, has generally limited functionality when it comes to matrix operations, so Numpy was used a lot for matrix manipulations. During the first implementation attempt, the solver gave mostly unintelligible error messages, so for the prototype SVM I actually used Matlab. Matlab's `quadprog` function had an identical specification to the Python solver I intended to use. Matrices in Matlab are a lot more straightforward to manipulate, which was also a good reason to first check the correctness of my matrices in Matlab.

Kernel Functions

Kernel functions are the central component of support vector machines, and its choice is highly dependent on the application. I have considered a variety of kernel functions and experimented with their combinations to see how the performance differs. Table 2.2 shows the kernel functions I have used. The

Name	$\boldsymbol{K(\vec{x}, \vec{y})}$
Linear	$ax \cdot y^T + c$
Radial Basis Function(RBF)	$\exp^{-\gamma\|\|x-y\|\|^2}$
Sigmoid	$\tanh(ax \cdot y^T + c)$
Polynomial	$a(x \cdot y^T)^d$
Weighted Sum	`a*RBF(x,y)+(1-a)*Linear(x,y)`
Product	`RBF(x,y)*RBF(x,y)`

Table 2.2: Kernel Functions

simplest kernel function is the linear kernel. It is simply a dot product of two vectors. Its generalized version is the polynomial kernel, which takes degree as a parameter. Intuitively, the polynomial kernel is a lot more flexible than linear, but the larger the degree, the less 'smooth' it becomes, so it might overfit the training data. The Radial Basis Function kernel is most popular in Support Vector Machines. We take $||x - y||$ to denote the Euclidean distance of the two feature vectors. The γ parameter is equivalent to $\frac{1}{2\sigma^2}$. The choice of γ has a major effect on the performance of the kernel and represents a tradeoff between over- and underfitting. To find the best parameter, cross validation is used. **TODO: CROSS VALIDATION** The sigmoid kernel is commonly used in neural networks and in combination with an SVM classifier forms a two layer perceptron network, where the scale parameter a is usually set to 1/N, where N is the number of dimensions (features)[4]. The Sum and Product kernels are both a combination of the linear and RBF kernels. In the Evaluation chapter we will discuss the relative performance of each of the kernels.

2.4 Parser

So far we have looked at the two separate blocks – the search engine and the machine learners. As a functional unit the parser must provide an interface between the two. The search engine simply retrieves pages in response to a query and the machine learner expects as its input a set of labelled features for training and unlabelled features for classification/regression. The Parser, therefore, must hide the nature of the data we are dealing with by translating it into the universal language of machine learners. Hence, the primary function of the Parser is to compute feature vectors for pages. However, it is easy to outsource search engine heuristics into the Parser, too, as they require page parsing.

To accomplish these multiple goals, I have taken an object oriented approach to the design of the module. What I refer to as the Parser is a few classes, which together perform a series of tasks related to page parsing. The module operates in two modes: rank and score and handles both classification and regression. The high level specification is that we create two objects: a `TestFeatureSetCollection` and a `LabelFeatureSetCollection` objects and they encapsulate all the data, so can be passed around to the machine learners, as well as evaluation and plotting modules. These objects each operate within their own directories, to keep training and testing sufficiently separate.

Both category and rank are treated as page features and hence are part of the `LabeledFeatureSet` class state. The `LabeledFeatureSetCollection` class generates training sets from queries, whereas the `TestFeatureSetCollection` class computes predicted category given an instance of a classifier. Both, however, need to generate feature sets, one for the Training data and the other for the Test data. This common functionality is embodied in their abstract base class, `FeatureSetCollection`. Figure 2.1 shows a UML class diagram illustrating the main structure of the module.

The rest of this section talks about the specifics of the implementation, in particular, the features used and how html parsing is done.

Rank mode. Conceptually a `FeatureSetCollection` is a dictionary of `LabeledFeatureSet` objects indexed by both path to page and query term. After initialisation, all the features are computed and 'filled up' per query term per page.

Score mode.

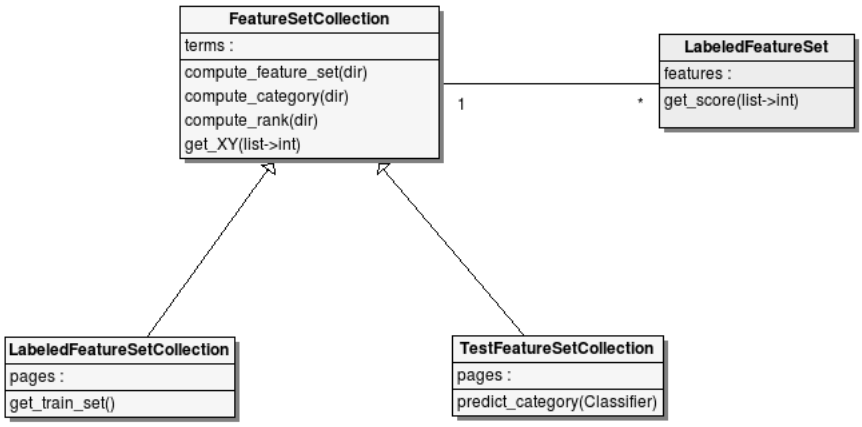


Figure 2.1: A UML class diagram describing operation of the Parser.

Features. The requirements analysis does not prescribe the use of any particular features. However, it states that both dynamic (query dependent) and static(query independent) features must be used. PageRank is a dynamic feature that has quite special place among others: it takes into account all the pages in the pool and reflects on the structural hierarchy of the web pages. The parser loads the pre-computed PageRank vector indexed by page name to extract the PageRank of any particular page.

An example of a dynamic feature used in the project is query term count: the number of times the query term occurs on the page. This is obtained directly from the html files using the BeautifulSoup library, just like for the link parsing in the crawler. The html is parsed into clean text and the count is computed on the string.

A related but different feature – stem count – is the number of times the stem of the word occurs in the text. This is obtained using the `PorterStemmer` module in Nltk library.

Boolean features are a slightly different variety of feature, so was worth putting in. An example of this is a presence of an image on the page.

The features have been added incrementally as the project progressed. All other features are similar to the ones described above and are obtained using the same methods.**TODO: enumerate features here**

2.5 Optimization

One characteristic peripheral requirement for the system implemented in this project is speed. Even though it is not our direct goal to produce efficient implementations, optimization could not be overlooked, because significant amount of time is spent processing large quantities of data: indexing, PageRank and feature set computations all must complete in 'feasible' time, i.e. in the final implementation the longest computation takes order of minutes and processes a few thousand pages. Various optimizations have been used to achieve this.

Both PageRank vector and the index are precomputed and kept in persistent storage. Incremental indexing feature allows us to edit parts of the index as opposed to recomputing the whole index from scratch. These precomputations provide certain speedups, but were not enough. Because the system is fairly complex and a lot of library code is used in places, it was hard to determine which code most affects the speed. 'Blind' attempts at optimization did not work well, which motivated the use of a profiling tool.

Profiling the first prototype of the complete system revealed a surprising fact: most time was spent in the library code parsing pages. To mitigate this issue I have tried using custom parsers instead. Apart from speed, robustness was another important consideration, as a failure of a parser increases compute time. Two of the most renowned python parsers are*html5lib* and *lxml*. Figure 2.2 below shows visual representation of time profiling of 3 different runs obtained using the *RunSnakeRun*, each exploiting a different parser. Despite *html5lib* being quoted as the most robust/lenient, *lxml* was sufficiently faster to be preferable.

(a) Default *(b) html5lib* *(c) lxml*

Figure 2.2: Visual profiles of three parsers from left to right: default python html parser, html5lib and lxml. The internal boxes are sized in proportion to the time spent in each function. In all profiles three distinct major compartments can be seen, the leftmost being the compartment of interest: time spent in parsing. Taking into consideration that the other modules are unaffected by changing the parser implementation, it can be observed that lxml (rightmost) is the fastest and html5lib is the slowest.

Another useful limitation was discovered due to profiling. The PageRank vector was loaded into memory every time a page was parsed: once for each page. This clearly is undesirable. I used caching to ensure that each of the two possible

PageRank vectors (corresponding to the test and train directories) is loaded from memory exactly once. This is achieved via a double singleton class, which loads the vectors lazily.

Chapter 3

Evaluation

Due to the incremental development measurements were taken at various stages as the implementations evolved. To avoid biased results Development corpus was used during development, which is disjoint from the real Test corpus and is of smaller size. The development corpus is further subdivided into the Training corpus and the Validation corpus to mirror the Training and the Test corpora, which are used to obtain all the results presented in this chapter. The evaluation was performed after the development of the whole system was complete and validated on the Development corpus.

3.1 Measurements

Talk about which statistical techniques are used for evaluation and explain why each was chosen and what were other alternatives considered.

3.1.1 Mean Squared Error

Second moment of the error, so incorporates both variance and bias. $\frac{1}{n}\sum_{i=1}^{n}(Actual - Predicted)^2$ Penalizes large errors heavily, but small errors diminish. Assuming Gaussian noise. But sensitive to outliers. MAE is less so. But MSE more visual and distinct representation of this particular data.

3.1.2 Tolerance intervals

Tolerance interval of the error mean is a good visualisation of how spread the error is. In the best case, the interval is tight. If the error is spread, the classifier

is making inconsistent mistakes. Particularly important to include these in our evaluation due to the presence of quantization error. The mean error of the ceiling and its spread is a good indicator of the mean and variance introduced by quantization.

3.1.3 Mean Reciprocal Rank

3.1.4 Classification and Quantization Error

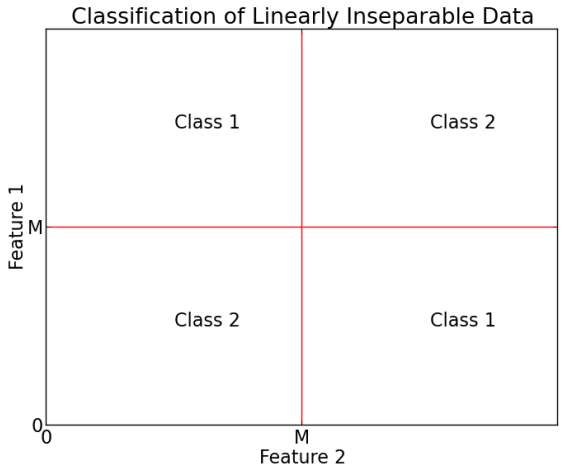
To compare classification and regression effectively, the error incurred in quantizing scores is tracked. In a two-class scenario there is one threshold score value that separates the two classes. This score value is computed as a median to have adequate class representation. When Bayes assigns a class to a feature set, it is perceived as assigning the mean value of the scores present in the class.

3.2 Linear and non-linear classification

Hypothesis: Naive Bayes performs better with linearly separable data. Naive Bayes is a linear classifier and, therefore, linear separability of data is a premise of successful classification. To test this hypothesis, we will construct two minimal two-dimensional examples as shown in figures 3.1 and 3.2 below. Example 3.1 cannot be solved by plotting a line to separate the classes. Bayes is expected to fail in this case. Example 3.2 illustrates a 'cut-off sum' function. Such a linear heuristic should be easy for Bayes to pick up on.

To evaluate the real Bayes behaviour the programs described above are ran with class 1 score equal to 1 and class 2 score equal to 50. Mean squared errors are computed for both cases. Figures 3.3 and 3.4 show the mean squared errors within their corresponding confidence intervals. Along with the actual Bayes classification error (denoted as 'Actual' on the y axes) the Baseline and Ceiling errors are plotted for comparison. The Baseline performance is evaluated by randomly assigning classes, whereas the Ceiling error illustrates the contribution of quantization error, inherent in classification of non-discrete scores.

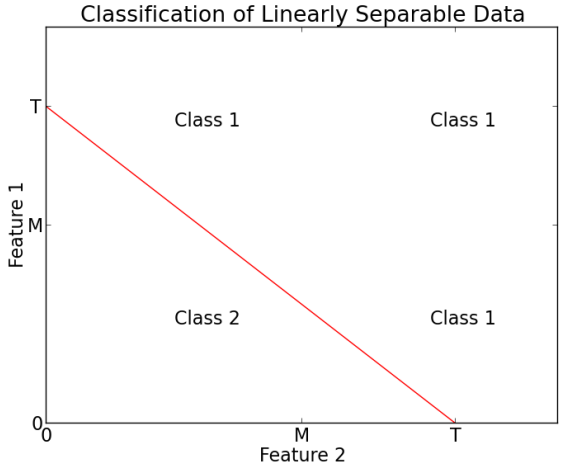
The mean squared error in the inseparable case is significantly larger: in fact, its confidence interval overlaps with the one of the error in random classification (the Baseline error). Similarly, the standard deviation of the error is similar in width to the Baseline. This is due to the fact that the classifier is making mistakes as frequently. In contrast, the separable data case shows a convincing improvement in classification error. However, the classifier is still not perfect. This might be a result of insufficient or unbalanced training data. In this




```
def score_inseparable:

    if f1 > M(f1):
        if f2 > M(f2):
            return class2
        else: return class1
    elif f2 > M(f2):
        return class1
    else: return class2
```

Figure 3.1: An example of inseparable data. M represents median points for corresponding features. The red lines visualize the boundaries between the classes, separating four distinct quadrants. The M value is chosen to be a median, so that adequate number of training examples was available for all quadrants.




```
def score_separable:

    if f1 +f2 > T:
        return class1
    else:
        return class2
```

Figure 3.2: An example of separable data. On the red line the sum of features is equal to the threshold value T. As before, T is chosen to be the median of the sum f1+f2 to supply equally many training examples on either side of the line.

particular case and in general, adding more training data improves the error up to a point when overfitting occurs. Nonetheless, it is rarely possible to gather a subset of training data that will have enough points very near to the class boundary to allow for perfect classification. Therefore, the result is consistent with the proposed hypothesis.

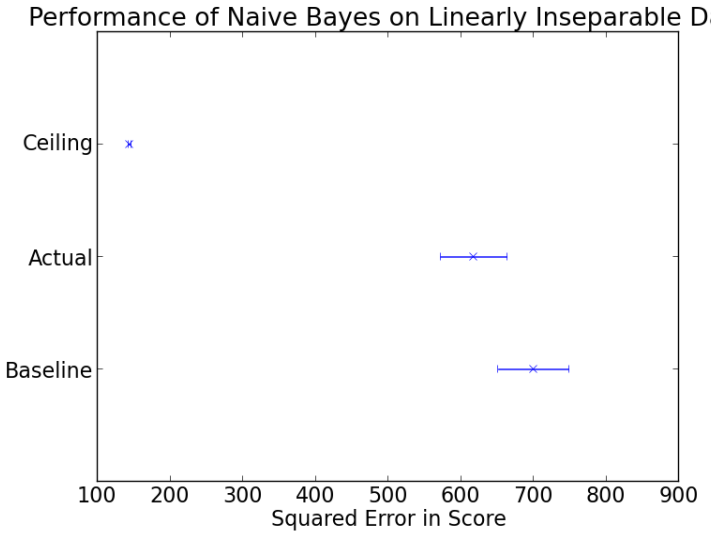


Figure 3.3: Classification of Inseparable Data: Evaluation. The mean error is near random classification and with wide spread.

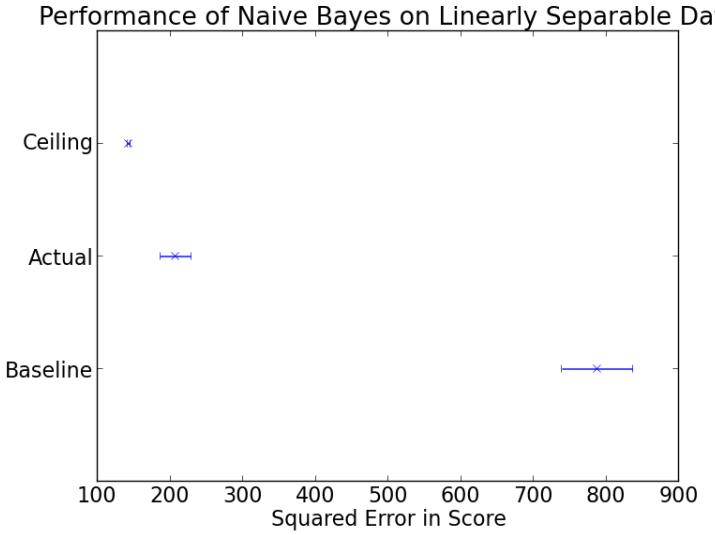


Figure 3.4: Classification of Separable Data: Evaluation. The mean error is distinctly better than random and approaches the goal performance.

SVM performs equally well regardless whether or not the data is linearly separable.

3.3 Hyperparameter tuning

Hypothesis: Bayes classifies better when fewer classes are present. It is intuitive that at a large number of classes, the more classes we introduce, the more precise the classifier has to be to perform equally well. However, taking into account the quantization error, we might expect non-linear behaviour: performance will improve due to the quantization error decreasing faster than

precision up to a certain point. This experiment aims to determine the number of classes minimizes the overall error of the classifier.

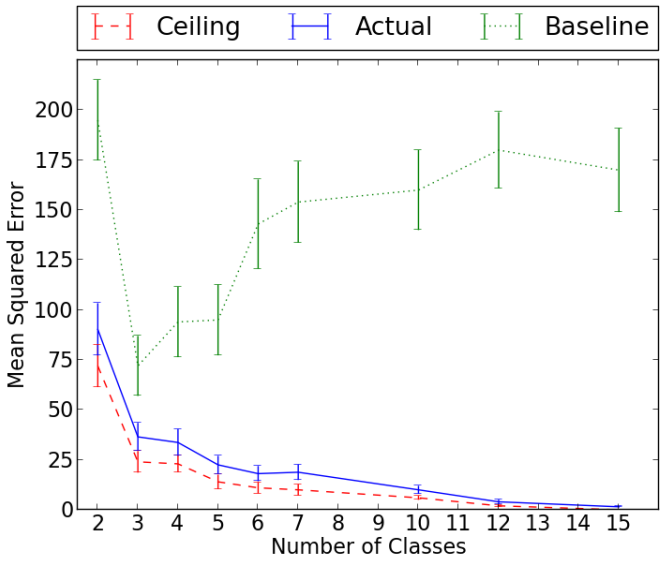
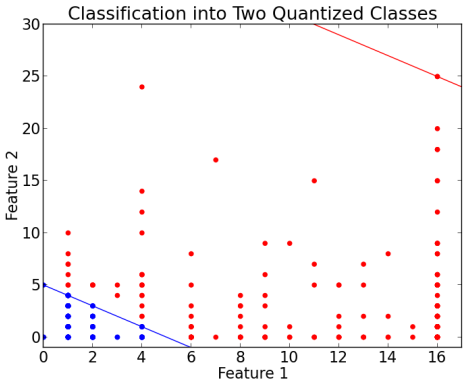
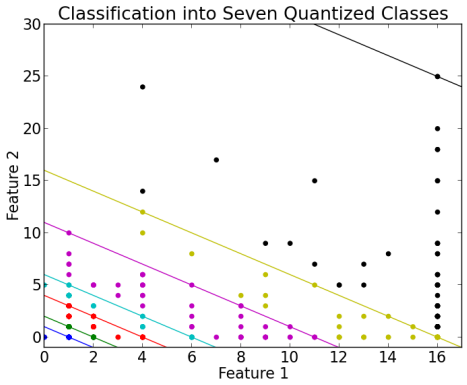


Figure 3.5: Error response to varying number of classes.



(a) TWO



(b) seven

SVM with linear kernel performs well for linear heuristics, but becomes unusable for non-linear ones. Grid Search + Cross validation Graphs.

Some kernels are more general than others.

3.4 Features

Bayes performance degrades rapidly as the number of features grows whereas SVM performance does not. So far we have considered two di-

mensional cases where only two features were used for convenience of visualization. Dimensionality curse is one of the well-known issues with classification: when new features are added, the size of the training set has to grow exponentially to cover all the new dimensions. Naive Bayes, however, lessens this problem a little due to the assumption of conditional independence. It is interesting to see how rapidly its performance degrades when more features are added while the training data is unchanged. It is important that all the features used are conditionally independent, as otherwise, the results would not be valid.

The number of classes is fixed at a value of 7, at which the performance is reasonable but still has some quantization error. This is done to prevent the effects of the extremes of performance. We are only interested in relative performance, so it should not matter how many classes are used, however, it is important that the number of classes used produces a significantly better result than the baseline, to make sure a random guess could not produce a good result.

The result for a few linearly separable functions is computed and the average errors are plotted in figure 3.6. The actual classification diverges from the ceiling when the number of features is about 6. The rapid growth of the random classification curve is explained by the curse of dimensionality: degrees of freedom grow exponentially. The exponential growth is plotted in black for comparison: initially the baseline growth is near exponential, but less steep subsequently. **TODO: why might that be?**

The ceiling and the actual curves are growing almost monotonously, as is expected. The flat regions at 5-6 and 7-8 are explained by the data specifics: the particular features added at those point appeared insignificant in the scores as they occurred infrequently and had small values.

It has proven hard to pick many conditionally independent features. Among the ones used were pagerank, image count, word count, frequency of search term occurrence, quality of html, price information availability and alike.

Both Bayes and SVM perform better when all the features used in training are also used in heuristics.

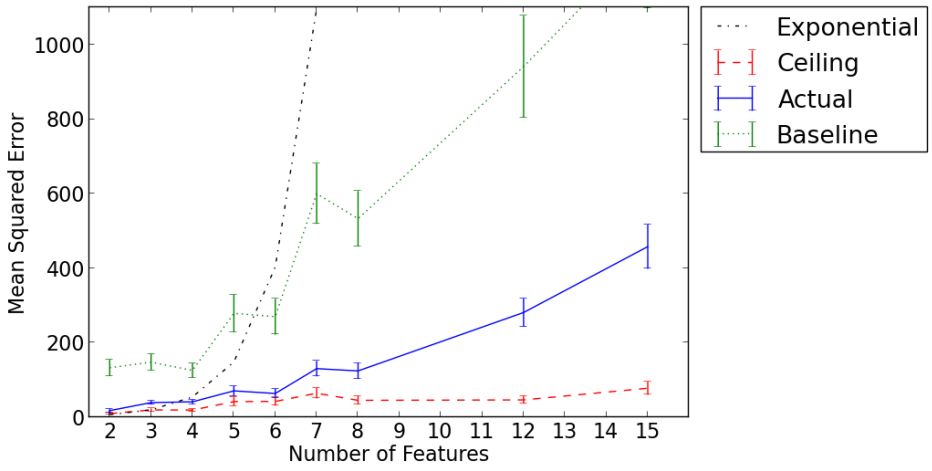


Figure 3.6: Error response with varying number of features.

Chapter 4

Conclusion

Bibliography

[1] Pedro Domingos. A few useful things to know about machine learning. University of Washington.

[2] Pedro Domingos. On the optimality of the simple bayesian classifier under zero-one loss. Kluwer Academic Publishers, 1997.

[3] Pedro Domingos and Michael Pazzani. Beyond independence: Conditions for the optimality of the simple bayesian classifier. pages 105–112, 1996.

[4] Hsuan-Tien Lin and Chih-Jen Lin. A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. National Taiwan University.

[5] Larry Page and Sergey Brin. The anatomy of a large-scale hypertextual web search engine. Stanford University, 1998.

[6] Larry Page and Sergey Brin. The pagerank citation ranking: Bringing order to the web. 1998.

[7] Jen-Wei Kuo Pu-Jen Cheng, Hsin-Min Wang. Learning to rank from bayesian decision inference. National Taiwan University.

[8] Harry Zhang and Jiang Su. Naive bayesian classifiers for ranking. pages 501–512, 2004.

Appendix A

Project Proposal