|  |
| --- |
| **SVM Project and Gender Classification** |
| **Josh Gleason and Rod Pickens** |
| **Programming Assignment 4** |
| **Computer Science 679 – Pattern Recognition, UNR, Dr. Bebis** |
| **Due: Monday May 4, 2015** |
| **Submitted: Monday Monday 4, 2015** |

Contents

[1 Abstract 3](#_Toc417774790)

[2 Technical Discussion 3](#_Toc417774791)

[2.1 Gender classification 4](#_Toc417774792)

[2.1.1 Eigenvectors 4](#_Toc417774793)

[2.1.2 Eigenfaces 4](#_Toc417774794)

[2.2 Machine Learning 5](#_Toc417774795)

[2.2.1 Ideal Learning 7](#_Toc417774796)

[2.2.2 Inductive Learning 7](#_Toc417774797)

[2.3 Support Vector Machines 13](#_Toc417774798)

[2.4 Kernel Functions 16](#_Toc417774799)

[3 Project 16](#_Toc417774800)

[3.1 Experiment 1: SVM 16](#_Toc417774801)

[3.1.1 16x20 Images 16](#_Toc417774802)

[3.1.2 48x60 Images 17](#_Toc417774803)

[3.2 Experiment 2: Bayesian 17](#_Toc417774804)

[3.2.1 Training Parameters 17](#_Toc417774805)

[3.2.2 Testing on 16x20 Images 17](#_Toc417774806)

[3.2.3 Testing on 48x60 Images 17](#_Toc417774807)

[4 Conclusion 17](#_Toc417774808)

[4.1 Part A: SVM 17](#_Toc417774809)

[4.2 Part B: Bayesian 17](#_Toc417774810)

[4.3 Comparison 17](#_Toc417774811)

[5 Contributors 17](#_Toc417774812)

# Abstract

This paper describes our research regarding the fourth class project for the Computer Science pattern recognition class CS 679 taught by Dr. Bebis who is Department Chair of the Computer Science Department at the University of Nevada in Reno, Nevada.

The primary topics included in the main body of this report are a description of the project, learning machines, induction theory, theory of support vector machines (SVM), theory of kernel functions, and classifier results. The paper also compares SVM with Bayes classifier.

We added a section to our project report concerning the general topic of the ***induction theory of learning***, and we did this to round our understanding of learning machines because the class lecture on SVMs introduced us to two inductive principles used in machine learning: empirical risk minimization versus structured risk minimization.[[1]](#footnote-1)

In researching these two terms (empirical risk minimization and structured risk minimization), we were amazed to learn about the abstraction in machine learning based on the inductive principle.

As we understand from our research, machine learning is based first on selecting an induction principle and then second based upon selecting a learning method. Given this new view of learning (statistical pattern recognition), we definitely appreciate the class lecture on SVMs because we learned about empirical risk minimization learning versus structured risk minimization learning.

Due to that lecture and our subsequent research to better understand the terminology, we learned a whole new abstraction of machine learning. We will discuss what we have learned in a tutorial section below on inductive principles and their associated learning methods.

# Technical Discussion

This project depends upon gender classification using as features eigenfaces which, in this project, are non-linearly mapped to higher dimensional abstract feature space where a hyper-plane is used as the class discriminant. The eigenfaces and associated classification theory were studied in the previous project and learned in the class on linear discriminant function. The theory of eigenfaces are discussed beginning in Section 2.1.

The theory for support vector machines is related to the general theory of inductive methods, and from this general theory, we derive support vector machines from the inductive method class known as structured risk minimization. The discussion of inductive methods begins in section 2.3 and concludes with the theory of support vector machines.

## Gender classification

### Eigenvectors

Eigenvectors are those vectors that are invariant in direction to the action of a matrix A on the vector as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

In the above, is a square nxn matrix, is an nx1 vector, and is a scalar. As mentioned above, the action of the matrix upon the vector is to scale, and to only scale, the vector and not to change the direction of the vector

### Eigenfaces

The eigenface approach taken and experimented in this paper is from the research of Turk and Pentland,[[2]](#footnote-2) and their work was motivated by the earlier works of Sirovich and Kirby who represented pictures using principal component analysis.[[3]](#footnote-3) Principal component analysis is a least squares approach for minimizing the error associated with a projection of the data onto a different basis.

Given a vector in an N dimensional space, it can be represented by a set of N orthogonal basis vectors as follows

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

The goal is to find an N x K transformation matrix U such that

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

Where is an Nx1 rasterized image

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

And the basis vectors the right side of the equation are the standard basis (natural or canonical) vectors of Euclidean space as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

Where is known as the Kronecker delta function

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

Yielding a set of basis vectors as follows

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

Moreover, the vector can be represented by a set of K orthogonal basis vectors *i*=1,…,K in a lower K (K<N) dimensional space, and each vector is an Nx1 vector. The lower K, again where K<N, dimensional space is defined as

|  |  |  |
| --- | --- | --- |
|  |  | (8) |

Principal component analysis selects the basis and coefficients to minimize the error

|  |  |  |
| --- | --- | --- |
|  |  | (9) |

## Machine Learning

Machine learning will be described using the same formalism as Cherkassky in that he specifics the learning as taking the inputs from both a generator of x and a system that processes x and outputs a y, and the learning machine receives these inputs x and y to produce an approximation to y. The output of the learning machine is an approximation to y because in the operational system we have no information on the true value of y, see Figure 1.



Figure : Learning Machine (Cherkassky[[4]](#footnote-4))

There are two aspects to the learning machine in the Figure 1: training and operation. Training the learning machine occurs when the two inputs x and y are received by the learning machine and a model of the generator and system are captured in the function f(x,y,w\*) that approximates the output of the system y. Operation of the learning machine occurs when only a single input x is input into the learning machine, and the learning machine outputs a prediction of the value y.

Mathematically, during training the learning machine builds a function f(x,y,w\*) that approximates the output from both the generator function g(x,w0) and the system function h(x,w0). The learning machine f(x,y,w\*) estimates a function f(x,w\*) from training data x and y with the purpose of using f(x,w\*) to predict h(x,w0) when h(x,y0) outputs are unavailable:

|  |  |  |
| --- | --- | --- |
|  |  | (10) |

The input x and y to the learning machine are independent and identically distributed (i.i.d.) pairs of data (*x,y*) distributed according to an unknown probability distribution function

|  |  |  |
| --- | --- | --- |
|  |  | (11) |

If the test samples have the same probability distribution as the training samples, then the learning machine should correctly label unknown input samples x with the appropriate label y.

The best function[[5]](#footnote-5) for learning should be that function that minimizes the expected risk (error)

|  |  |  |
| --- | --- | --- |
|  |  | (12) |
|  |  |  |
|  |  |  |

The function is the loss function and P(x,y) is the known joint probability density function of the input x and y. Assuming a 0/1 loss function, then the expected risk is the minimum error function.

### Ideal Learning

The ideal learning machine is one such that the researcher is completely aware of the model of the probability distribution for the data x, the parameters describing the distribution, the system function h(x,w), and the cost function. With this a-priori knowledge, a researcher can design an optimal learning machine that minimizes the expected risk.

|  |  |  |
| --- | --- | --- |
|  |  | (13) |

### Inductive Learning

When the conditions for ideal learning are unavailable, e.g. the probability distribution is unknown, then a researcher will be required to select from one of a number of inductive learning principles with which to base his machine learning engine. Once an inductive principle is selected, then the researcher will have to select a particular learning method from that inductive principle to construct the learning machine.

To illustrate with an example, assume that the researcher has no a-priori knowledge of the probability distribution for x and y. Given this, the researcher can proceed in the design of the learning machine basing on using a simple induction principle. A particularly simple induction principle is use is to base the design upon an approximation to the expected risk. Defining the loss function as the 0/1 or symmetric loss, the empirical risk can be defined as the sample mean of the loss associated with the outcomes of the experiment as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (14) |

Note that the empirical risk should asymptotically approach the expected risk.

According to Cherkassky,[[6]](#footnote-6) there are a number of inductive principles to learning:

1. Empirical Risk Minimization
2. Structural Risk Minimization
3. Penalization
4. Bayesian Inference
5. Minimum Description Length

The author primarily focuses on the four inductive principles of Penalization, Bayesian Inference, Structured Risk Minimization, and Minimum Description Length, but he gives a short description of empirical risk minimization (to be discussed later).



Figure : Inductive Principles in Learning

Cherkassky repeatedly highlights that there is an important distinction between inductive principles and learning methods. For a given inductive principle, many different learning methods (even infinitely many) can be used to execute the inductive principle. These different learning methods correspond to different classes of approximation functions (the decision rule g(x)), parameters, and the optimization technique (how the g(x) is found). The learning methods are also known as constructive implementations of an induction principle as shown in Figure 3.

Figure : Learning methods for a given inductive principle

An inductive method is concerned with a general approach to how to use data for learning, and the learning method that one uses from a particular inductive method is the implementation that gives the estimate, or what to do with the data.

Cherkassky segregates inductive principles into classical and adaptive, and he distinguishes the two inductive principles based on the amount of information available prior to learning. In other words, the two classes are distinguished by the amount of information assumed as valid prior to learning, Figure 1.

The classical inductive methods (e.g. empirical risk minimization) assume strong a-priori knowledge of the probability distributions whereas the adaptive inductive methods assume weak a-priori knowledge of the probability distributions.

To illustrate the above concepts, the author Cherkassky illustrates empirical risk minimization (ERM) along with the various learning methods that can be implemented under the rubic of ERM such as maximum likelihood, linear regression, polynomial regression, and fixed-topology neural networks.

In ERM, the model is given, e.g. Gaussian density function, and then the parameters for that Gaussian distribution are found, e.g. mean and standard deviation. The parameters can be found using a method such as maximum likelihood. ERM works well when the number of training samples is large relative to the model complexity (number of free parameters such as mean and standard deviation).

For example, if only one free parameter is required to be estimated, e.g. the sample mean, and if many samples are available for estimating the sample mean, then maximum likelihood is a good estimator for the mean (assuming that assumed probability density model is the correct density model). If however, the number of free parameters (mean, covariance matrix) that must be estimated is large relative to the number of training samples, then the ERM methods performs poorly. Other factors can enter into the quality of the learning machine. The researcher might want to learn the mean, but it could be the case that mean is an insufficient statistic (not well estimated by the sample mean) for the model.

#### Empirical Risk Minimization

Empirical risk minimization is the inductive approach associated with the classical statistical approaches of “parametric estimation” and “classification” as developed by R.A. Fisher (1890-1962).[[7]](#footnote-7)

R.A. Fisher was trained as a mathematician, and as a student, he had not formally studied statistics. He became involved in statistics during his first job with the city of London and in a later job working at the Rothamsted Experimental Station (one of the oldest agriculture research institutes in the world) in Harpenden, Herfordshire, England. He started working at the experimental station in the year 1919, and while at the station he performed agriculture crop research. During this time, Fisher used and developed modern statistical concepts to support his analysis in crop variation. His research also supported his personal passion for studies in eugenics.

During this time period at the experimental station, Fisher became especially involved in statistical analysis and the design of experiments after reading William Gosset’s paper on the Student’s t- finite sampling distribution. Later, he became interested in curve fitting after reading Karl Pearson’s method of moments for curve fitting, which Fisher questioned as an appropriate method, and Fisher then recast Pearson’s curve fitting approach into the method maximum likelihood. Fisher published his new approach which evidently created friction between himself and Pearson.[[8]](#footnote-8) During Fisher’s lifetime, he nearly single-handedly created modern empirical statistical theory, which falls under the inductive method of empirical risk minimization.

Under the rubic of empirical risk minimization, we learn that maximum likelihood is a specific “constructive method” or “learning method” used to estimate the parameters given that the researcher can assume the correct probability density.[[9]](#footnote-9) For maximum likelihood, the researcher assumes a probability distribution for the data under question for which the distribution parameters are to be estimated, and the parameters are estimated based on “empirically” collected data (empirical in that the data is collected from an experiment).

In more detail, maximum likelihood is associated with a particular learning method that minimizes the Kullback-Leiber divergence metric (as used in information theory).[[10]](#footnote-10)

#### Structural Risk Minimization

Structural risk minimization was first presented in a paper by Vladimir Vapnik and Alexey Chervonenkis[[11]](#footnote-11) in 1974 as a method of optimal learning that combines “empirical risk minimization” with an added constraint to minimize the complexity of the model (that is used to make decisions.[[12]](#footnote-12) In essence, the structural risk principle favors learning machines that have less complex models (e.g. the machine is based on low order polynomials rather than high order polynomials).

The author’s formulated a theory of statistical learning with the goal of generalized learning, and towards this goal the author’s determined to control both the empirical risk and the generalization ability of a learning machine by two factors: error-rate on training data and the capacity or confidence of the learning machine[[13]](#footnote-13):

|  |  |  |
| --- | --- | --- |
|  |  | (15) |

The confidence is related to the Vapnik-Chervonenkis (VC) dimension (h) of the decision function class (how complex the decision function is) and the number of training samples (n) as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (16) |

The confidence interval holds with probability 1-, and note that for this to hold n > h.

The general intuition behind the statistical learning theory is that a researcher can design a learning machine with a high complexity decision boundary (high VC dimension) such that the empirical training error goes to zero. This is generally not desirable. In the case where the training errors go to zero, it is often the case that the confidence could be quite large while the learning machine predicts poorly; in other words, the classifier error >> expected risk.

At the other extreme, the researcher could lower the confidence interval to nearly zero at the risk of increasing the empirical error (too much generalization) which gives a learning machine poor performance; in other words, the error >> expected risk.[[14]](#footnote-14)

Vapnik and Chervnonkis recommend designing the learning machine so as to find a balance between confidence and empirical risk.

From the statistical learning theory developed by Vapnik and Chervnonkis, the induction principle of structural risk minimization was formulated and leads to a class of learning machines known as support vector machines (SVMs) which are linear classifiers. SVMs are often combined with non-linear functions that map the inputs x into a higher dimensional feature space to improve classification. In this higher dimensional space , the learning machine f(x,w) can build a low complexity decision boundary, such as a hyper-plane, such that the classifier error is minimized and the generalization ability of the classifier is maximized.[[15]](#footnote-15)

#### Penalization Induction Principle

The risk[[16]](#footnote-16) used in the inductive principle of penalization is determined as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (17) |

The empirical risk denotes training error, and the penalty function is a non-negative functional associated with each possible estimate of f(x,w). The parameter weights the penalty relative to the empirical risk in decision making. The genetic algorithm discussed in class appears to be a machine learning method associated with penalized induction.

#### Minimum Description Length Induction Principle

This principle[[17]](#footnote-17) contrasts with all other methods in that the approach regards models as codes, as in information theory. The approach is to encode the training data such that the code length represents the data in a way that generalizes the learning machine.

#### Bayes Induction

Bayes induction principle[[18]](#footnote-18) is based on having available a-priori probability information along with a-priori defined likelihood functions. Given the a-priori probabilities, a classifier is defined that reduces empirical risk. Assuming that the loss function is the 0/1 loss function, then the Bayesian induction produces a learning machine based on Bayes’ principle

|  |  |  |
| --- | --- | --- |
|  |  | (18) |

With the 0/1 loss function, the Bayes classifier minimizes the expected risk and minimizes the error. The Bayes decision rule then becomes defined as

|  |  |  |
| --- | --- | --- |
|  |  | (19) |

As an interesting side note, Cherkassky states that the penalization inductive method has been formulated in terms of Bayes induction by some researchers given that the penalty and the priors can be considered similar, but Cherkassky concludes otherwise because the penalty is not based on a prior probability distribution.[[19]](#footnote-19)

## Support Vector Machines

The support vector machine (SVM) is a learning machine or constructive method that belongs to the structural risk minimization inductive principle, as described above. The SVM is primarily a two-class classifier that can be extended to multiple-classes. The optimization criterion is to maximize the margin of separation between the features of the classes which ensures that the VC dimension is minimized along with the empirical risk. Training a SVM is equivalent to solving a quadratic programming problem with linear constraints.[[20]](#footnote-20)

To describe the SVM, the inventors assume classification without training error (that which minimizes the empirical error) because they assume complete class separability. Given this assumption, the following linear functional defines a hyper-plane between classes that operates with no error (assuming the classes are separable):

|  |  |  |
| --- | --- | --- |
|  |  | (20) |

Where and is a feature vector. The classifier derived from the above functional is simply

|  |  |  |
| --- | --- | --- |
|  |  | (21) |

The learning machine defines both learning the vector w (the vector perpendicular to the hyperplane separating the classes) and the parameter b (the position of the hyperplane). With properly defined w and b, a support vector learning machine maximizes the distance between the nearest neighbors (support vectors) of the classes with the decision boundary. The boundary of the decision surface, the hyperplane, is given as

|  |  |  |
| --- | --- | --- |
|  |  | (22) |

And as shown in Figure 4



Figure : Hyperplane wx + b (defined by vector w) and margin

The distance between the classes as defined by the boundary can be determine from two exemplars and that lie on the boundaries: and . The distance between the lines is determine as

|  |  |  |
| --- | --- | --- |
|  |  | (23) |

The maximization of the distance across the margin, which in turn reduces the VP dimension, is done by quadratic programming to minimize as follows

|  |  |  |
| --- | --- | --- |
|  |  | (24) |

We are unable to solve this minimization directly, so we use Langrangian multipliers to find w and b, as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (25) |

The solution is rather straight forward, but prior to solving this optimization, we introduce kernel functions to map the feature vector x non-linearly to a higher dimensional space to ensure class separability (as assumed above). The kernel function comes to us via a non-linear function that directly maps x into a higher (possibly much higher) dimensional feature space in which the classes become linearly separable. The function is defined as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (26) |

Given this function, one develops the learning algorithm in the space rather than original space with training samples now defined as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (27) |

Where . We now substitute for x into the above Langrangian

|  |  |  |
| --- | --- | --- |
|  |  | (28) |

With this and to optimize the learning machine, we maximize with respect to w and b, and then we minimize with respect to the Langrangian multipliers .[[21]](#footnote-21)

Taking the partials of the Langrangian and setting to zero to find the extremes, we have

|  |  |  |
| --- | --- | --- |
|  |  | (29) |

And

|  |  |  |
| --- | --- | --- |
|  |  | (30) |

Solving the first partial gives us

|  |  |  |
| --- | --- | --- |
|  |  | (31) |

Which reduces to

|  |  |  |
| --- | --- | --- |
|  |  | (32) |

Solving the second partial gives us

|  |  |  |
| --- | --- | --- |
|  |  | (33) |

Which reduces to

|  |  |  |
| --- | --- | --- |
|  |  | (34) |

The latter is an amazing result because it gives us convolution. We believe this because the optimization produces a learning machine that involves producing the outcome from a sum of dot products, which in turn is a convolution of the feature vectors in the higher dimensional space.

Substituting w into the Langrangian and maximizing based on the Langrangian multipliers we get the following quadratic optimization problem

|  |  |  |
| --- | --- | --- |
|  |  | (35) |

Substituting a kernel function for the non-linear mapping functions , we get the quadratic optimization problem:

|  |  |  |
| --- | --- | --- |
|  |  | (36) |

With the following linear constraints

|  |  |  |
| --- | --- | --- |
|  |  | (37) |

And

|  |  |  |
| --- | --- | --- |
|  |  | (38) |

In matrix notation, the above simplifies to

|  |  |  |
| --- | --- | --- |
|  |  | (36) |

This clearly shows the fact that the optimization is on a quadratic surface with the following constraints

|  |  |  |
| --- | --- | --- |
|  |  | (36) |

Here the individual entries in the matrix are defined as

## Kernel Functions

# Project

This project consists of several experiments to compare various versions of the eigenface recognition algorithm on different data sets.

## Experiment 1: SVM

### 16x20 Images

#### Polynomial Kernels

#### RBF Kernels

### 48x60 Images

#### Polynomial Kernels

#### RBF Kernels

## Experiment 2: Bayesian

### Training Parameters

### Testing on 16x20 Images

### Testing on 48x60 Images

# Conclusion

## Part A: SVM

## Part B: Bayesian

## Comparison

# Contributors

1. Cherkassky, Vladimir, “Inductive Principles for Learning from Data,” http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.26.7820&rep=rep1&type=pdf [↑](#footnote-ref-1)
2. M. Turk and A. Pentland, “Eigenfaces for Recognition,” Journal of Cognitive Neuroscience, vol. 3, no. 1 pp. 71-86, 1991. [↑](#footnote-ref-2)
3. L Sirovich and M Kirby, “Low-dimensional procedure for the characterization of human faces,” Journal of the Optical Society of America A, 4(3), 519-524. [↑](#footnote-ref-3)
4. Cherkassky, Vladimir, “Inductive Principles for Learning from Data,”page 2. [↑](#footnote-ref-4)
5. Klaus-Robert Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernard Scholkoph, “An Introduction to Kernel-Based Learning Algorithms,” IEEE Transactions on Neural Networkds, Vol. 12, No. 2, March 2001. [↑](#footnote-ref-5)
6. Ibid, page 5. (Cherkassky, Vladimir, “Inductive Principles for Learning from Data”) [↑](#footnote-ref-6)
7. http://en.wikipedia.org/wiki/Ronald\_Fisher [↑](#footnote-ref-7)
8. Spanos, Arias, “R. A. Fisher: how an outsider revolutionized statistics,” http://errorstatistics.com/2014/02/17/r-a-fisher-how-an-outsider-revolutionized-statistics-2/ [↑](#footnote-ref-8)
9. Ibid, page 8. (Cherkassky, Vladimir, “Inductive Principles for Learning from Data” ) [↑](#footnote-ref-9)
10. Kevin Swersky, “Inductive Principles for Learning Restricted Boltzmann Machines,” Master’s Thesis, University of British Columbian, August 2010. [↑](#footnote-ref-10)
11. http://en.wikipedia.org/wiki/Structural\_risk\_minimization [↑](#footnote-ref-11)
12. Bebis, George, Lecture on “Support Vector Machines,” page 7, 2015. http://www.cse.unr.edu/~bebis/CS479/ [↑](#footnote-ref-12)
13. Corinna Cortes and Vladimir Vapnik, “Support-Vector Networks,” Machine Learning, 20, pages 273-297, 1995. [↑](#footnote-ref-13)
14. Klaus-Robert Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernard Scholkoph, “An Introduction to Kernel-Based Learning Algorithms,” IEEE Transactions on Neural Networkds, Vol. 12, No. 2, March 2001. [↑](#footnote-ref-14)
15. Vladimir Vapnik, “The Nature of Statistical Leaning Theory,” Springer-Verlag, New York, 1995. [↑](#footnote-ref-15)
16. Cherkassky, Vladimir, “Inductive Principles for Learning from Data,”page 8. [↑](#footnote-ref-16)
17. Ibid, page 13. (Cherkassky, Vladimir, “Inductive Principles for Learning from Data.”) [↑](#footnote-ref-17)
18. Ibid, page 11. (Cherkassky, Vladimir, “Inductive Principles for Learning from Data.”) [↑](#footnote-ref-18)
19. Ibid, page 13. (Cherkassky, Vladimir, “Inductive Principles for Learning from Data”) [↑](#footnote-ref-19)
20. Bebis, George, Lecture on “Support Vector Machines,” page 7, 2015. http://www.cse.unr.edu/~bebis/CS479/ [↑](#footnote-ref-20)
21. Klaus-Robert Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernard Scholkoph, “An Introduction to Kernel-Based Learning Algorithms,” IEEE Transactions on Neural Networkds, Vol. 12, No. 2, March 2001. [↑](#footnote-ref-21)