Physica A 437 (2015) 101–108

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Application of the method of maximum entropy in the mean to classification problems

Henryk Gzyl [a,*], Enrique ter Horst [b], German Molina [c]

[a] Centro de Finanzas, IESA, Venezuela
[b] Colegio de Estudios Superiores de Administración-CESA, Colombia
[c] Idalion Capital Group, USA

HIGHLIGHTS

- The classification problem appears in a large variety of fields.
- It can be regarded to be linear inverse problem with convex constraints and data in sets.
- An important version of the problem appears in the context of credit scoring.
- The problem can be solved efficiently with the method of maximum entropy in the mean.
- The solution obtained with MEM competes favorably when compared to standard methods like SVM and kNN.

ARTICLE INFO

Article history:
Received 13 January 2015
Received in revised form 12 April 2015
Available online 30 May 2015

Keywords:
Maximum entropy in the mean
Classification problems
Credit scoring
Linear inverse problems

ABSTRACT

In this note we propose an application of the method of maximum entropy in the mean to solve a class of inverse problems comprising classification problems and feasibility problems appearing in optimization. Such problems may be thought of as linear inverse problems with convex constraints imposed on the solution as well as on the data. The method of maximum entropy in the mean proves to be a very useful tool to deal with this type of problems.

1. Introduction and preliminaries

1.1. Motivational problems

The statement of the problem that we are going to be dealing with and a proposed solution, can be traced back to Fisher [1] as a problem in statistical classification. It is a problem that appears in many contexts, like pattern recognition, internet search engines, micro-array classification, computer vision, geostatistics, etc. Actually, in [2] there are many interesting theoretical and applied aspects of the classification problem are treated, as well as a long list of references to applications. We chose to motivate it as a problem in credit scoring for its intuitive appeal, see Ref. [3] for details.

Suppose that there are two populations: the "good" customers and the "bad" customers. The goodness or badness of an individual is described by an N-dimensional vector $\mathbf{a}$ of characteristics that are related to his probability of default. Suppose now that our data consists of two collections of characteristic vectors $\{\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_K^{(1)}\}$ and $\{\mathbf{a}_1^{(2)}, \ldots, \mathbf{a}_K^{(2)}\}$, corresponding to

* Corresponding author.
E-mail addresses: henryk.gzyl@iesa.edu.ve (H. Gzyl), enrique.terhorst@cesa.edu.co (E. ter Horst), german@germanmolina.com (G. Molina).

http://dx.doi.org/10.1016/j.physa.2015.05.105
0378-4371/

each type of individual, which we can suppose without loss of generality, to be of the same size. Suppose that we want to develop a "quick" method to separate these two populations into groups in such a way that when a new set of characteristics is given, we may easily tell whether the individual comes from the "good" or the "bad" guys. For example, that we want to split $\mathbb{R}^N$ into two disjoint subsets, in such a way that it is easy to decide to which of them a new vector of characteristics belongs to.

For example, an easy way to separate the population into groups, consists in determining a $K-1$ subspace (a hyper-plane passing through the origin in $\mathbb{R}^K$) such that the good guys are characterized by points on one side of it and the bad guys by points on the other side. That is, if $\mathbf{w}$ is an N-vector normal to the hyper-plane, we have $\langle \mathbf{a}_i^{(1)}, \mathbf{w}\rangle > 0$ for all $i = 1, \ldots, K$, while $\langle \mathbf{a}_i^{(2)}, \mathbf{w}\rangle < 0$ for all $i = 1, \ldots, K$, where we use the standard notation $\langle \mathbf{u}, \mathbf{v}\rangle$, for the Euclidean scalar product of the vectors $\mathbf{u}$ and $\mathbf{v}$. We now assemble the $2K \times N$ matrix $\mathbf{A}^0$ as follows:

$$\mathbf{A}^0_{i,j} = \begin{cases} A^0_{i,j} = \mathbf{a}_i^{(1)}(j) & 1 \le i \le K \; 1 \le j \le N. \\ A^0_{i,j} = -\mathbf{a}_i^{(2)}(j) & K+1 \le i \le 2K \; 1 \le j \le N. \end{cases}$$

Set $M = 2K$ for typographical convenience, and our problem can be stated as

$$\text{Determine a vector } \mathbf{w} \text{ such that } \mathbf{A}^0\mathbf{w} \in \mathbb{R}^M_{++} \tag{1}$$

where $\mathbb{R}^M_{++}$ denotes the orthant of strictly positive vectors. Observe now, that the classification scheme can be extended a bit to include possible misclassification. We may want the good guys to satisfy $\langle \mathbf{a}^{(1)}, \mathbf{w}\rangle > -\ell$, while the bad guys satisfy $\langle \mathbf{a}^{(2)}, \mathbf{w}\rangle < \ell$, for some positive ℓ to be determined. If we introduce the $2K$ dimensional vector $\mathbf{u}$ with all of its components equal to 1, and call $\mathbf{x}$ the column vector $(\mathbf{w}, \ell)^t$ then problem (1) may be rephrased as

$$\text{Determine a vector } \mathbf{x} \text{ such that } \mathbf{A}\mathbf{x} \in \mathbb{R}^M_{++} \tag{2}$$

where now $\mathbf{A} = [\mathbf{A}^0\mathbf{u}]$, that is we augment $\mathbf{A}^0$ by juxtaposing an M-dimensional column vector of ones next to it.

$$\text{Let } \mathbf{x}^* = \begin{pmatrix} \mathbf{w}^* \\ \ell^* \end{pmatrix} \text{ denote a solution to (2) and set } \mathbf{n} = \mathbf{w}^*/\|\mathbf{w}^*\| \; b = \ell^*/\|\mathbf{w}\|^*$$

which also solves (2). Now, define the hyper-planes

$$\begin{aligned} H_+ &= \{\mathbf{y} \in \mathbb{R}^N \,|\, \langle \mathbf{y}, \mathbf{n}\rangle = b\} \\ H_- &= \{\mathbf{y} \in \mathbb{R}^N \,|\, \langle \mathbf{y}, \mathbf{n}\rangle = -b\}. \end{aligned} \tag{3}$$

These two hyper-planes determine three regions which we now describe as

$$\begin{aligned} G &= \{\mathbf{y} \in \mathbb{R}^N \,|\, \langle \mathbf{y}, \mathbf{n}\rangle + b > 0\} \\ B &= \{\mathbf{y} \in \mathbb{R}^N \,|\, \langle \mathbf{y}, \mathbf{n}\rangle - b < 0\} \\ U &= \{\mathbf{y} \in \mathbb{R}^N \,|\, -b \le \langle \mathbf{y}, \mathbf{n}\rangle \le b\}. \end{aligned} \tag{4}$$

Here, the sets G, B, U contain, respectively, the good, bad and hard to decide upon points. In the context of the credit scoring example, G contains the characteristics of the individuals that will be granted credit on the basis of their characteristics, B those that will be refused credit and U those which will be the subject of further examination or requirements.

Problem (1) or its extension (2) also appear as the initial step (determination of feasibility sets) in almost every linear programming problem The original classification problem also appears as a binary classification problem in the perceptron theory of leaning. Here the vectors $\{\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_K^1\}$ and $\{\mathbf{a}_1^{(2)}, \ldots, \mathbf{a}_K^2\}$, are called the training set and the problem of finding the vector $\mathbf{w}$, that is solving (1) is called the training problem for the perceptron. This problem has branched out in many directions and has been attacked by methods in statistical physics. For a very short list of papers on this, see Refs. [4–7].

There is a collection of methods/algorithms/procedures to tackle those problems. To name a few, consider support vector machine (SVM), k-nearest neighbor (KNN), or linear discriminant analysis. Besides the references cited above, consider the paper by Soheili and Peña [8], in which the authors examine some speed of convergence results for a combination of algebraic and analytic methods to obtain a solution to a problem like (1). Their notational conventions differ slightly from ours.

It is the purpose of this note to propose and develop one more method of solution, based on the method of maximum entropy in the mean (MEM), which is a general method to solve linear problems with convex constraints. We shall recall the basics about it below. Here, let us only add that MEM uses the standard method of maximum entropy developed to understand equilibrium statistical mechanics, as a stepping stone to solve linear inverse problems with convex constraints.

1.2. Mathematical preliminaries

Here we shall state our original problems in a somewhat more general framework. Actually the notational generality allows to set up and solve a larger class of similar problems with the same effort.