Estimating Car Insurance Premia: 
a Case Study in High-Dimensional Data Inference

Nicolas Chapados, Yoshua Bengio, Pascal Vincent, Joumana Ghosn, Charles Dugas, Ichiro Takeuchi, Linyan Meng
University of Montreal, dept. IRO, CP 6128, Succ. Centre-Ville, Montreal, QC, Canada, H3C3J7
{chapados, bengioy, vincentp, ghosn, dugas, takeuchi, mengl}@iro.umontreal.ca

Abstract

Estimating insurance premia from data is a difficult regression problem for several reasons: the large number of variables, many of which are discrete, and the very peculiar shape of the noise distribution, asymmetric with fat tails, with a large majority zeros and a few unreliable and very large values. We compare several machine learning methods for estimating insurance premia, and test them on a large data base of car insurance policies. We find that function approximation methods that do not optimize a squared loss, like Support Vector Machines regression, do not work well in this context. Compared methods include decision trees and generalized linear models. The best results are obtained with a mixture of experts, which better identifies the least and most risky contracts, and allows to reduce the median premium by charging more to the most risky customers.

1 Introduction

The main mathematical problem faced by actuaries is that of estimating how much each insurance contract is expected to cost. This conditional expected claim amount is called the pure premium and it is the basis of the gross premium charged to the insured. This expected value is conditioned on information available about the insured and about the contract, which we call input profile here. This regression problem is difficult for several reasons: large number of examples, large number variables (most of which are discrete and multi-valued), non-stationarity of the distribution, and a conditional distribution of the dependent variable which is very different from those usually encountered in typical applications of machine learning and function approximation. This distribution has a mass at zero: the vast majority of the insurance contracts do not yield any claim. This distribution is also strongly asymmetric and it has fat tails (on one side only, corresponding to the large claims).

In this paper we study and compare several learning algorithms along with methods traditionally used by actuaries for setting insurance premia. The study is performed on a large database of automobile insurance policies. The methods that were tried
are the following: the constant (unconditional) predictor as a benchmark, linear regression, generalized linear models (McCullagh and Nelder, 1989), decision tree models (CHAID (Kass, 1980)), support vector machine regression (Vapnik, 1998), multi-layer neural networks, mixtures of neural network experts, and the current premium structure of the insurance company.

In a variety of practical applications, we often find data distributions with an asymmetric heavy tail extending out towards more positive values. Modeling data with such an asymmetric heavy-tail distribution is essentially difficult because outliers, which are sampled from the tail of the distribution, have a strong influence on parameter estimation. When the distribution is symmetric (around the mean), the problems caused by outliers can be reduced using robust estimation techniques (Huber, 1982; F.R. Hampel et al., 1986; Roussieuv and Leroy, 1987) which basically intend to ignore or downweight outliers. Note that these techniques do not work for an asymmetric distribution: most outliers are on the same side of the mean, so downweighting them introduces a strong bias on its estimation: the conditional expectation would be systematically underestimated.

There is another statistical difficulty, due to the large number of variables (mostly discrete) and the fact that many interactions exist between them. Thus the traditional actuarial methods based on tabulating average claim amounts for combinations of values are quickly hurt by the curse of dimensionality, unless they make hurtful independence assumptions (Bailey and Simon, 1960). Finally, there is a computational difficulty: we had access to a large database of $\approx 8 \times 10^6$ examples, and the training effort and numerical stability of some algorithms can be burdensome for such a large number of training examples.

This paper is organized as follows: we start by describing the mathematical criteria underlying insurance premia estimation (section 2), followed by a brief review of the learning algorithms that we consider in this study, including our best-performing mixture of positive-output neural networks (section 3). We then highlight our most important experimental results (section 4), and in view of them conclude with an examination of the prospects for applying statistical learning algorithms to insurance modeling (section 5).

2 Mathematical Objectives

The first goal of insurance premia modeling is to estimate the expected claim amount for a given insurance contract for a future one-year period (here we consider that the amount is 0 when no claim is filed). Let $X \in \mathbb{R}^m$ denote the customer and contract input profile, a vector representing all the information known about the customer and the proposed insurance policy before the beginning of the contract. Let $A \in \mathbb{R}^+$ denote the amount that the customer claims during the contract period; we shall assume that $A$ is non-negative. Our objective is to estimate this claim amount, which is the pure premium $p_{\text{pure}}$ of a given contract $x$:\footnote{The pure premium is distinguished from the premium actually charged to the customer, which must account for the risk remaining with the insurer, the administrative overhead, desired profit, and other business costs.}

$$p_{\text{pure}}(x) = E[A|X = x].$$

The Precision Criterion. In practice, of course, we have no direct access to the quantity (1), which we must estimate. One possible criterion is to seek the most precise estimator, which minimizes the mean-squared error (MSE) over a data set $D = \{(x_i, a_i)\}_{i=1}^L$. Let $\mathcal{P} = \{p(\cdot; \theta)\}$ be a function class parametrized by the
parameter vector $\theta$. The MSE criterion produces the most precise function (on average) within the class, as measured with respect to $D$:

$$\theta^* = \arg\min_{\theta} \frac{1}{L} \sum_{i=1}^{L} (p(x_i; \theta) - a_i)^2. \quad (2)$$

Is it an appropriate criterion and why? First one should note that if $p_1$ and $p_2$ are two estimators of $E[A|X]$, then the MSE criterion is a good indication of how close they are to $E[A|X]$, since by the law of iterated expectations,

$$E[(p_1 (X) - A)^2] - E[(p_2 (X) - A)^2] = E[(p_1 (X) - E[A|X])^2] - E[(p_2 (X) - E[A|X])^2],$$

and of course the expected MSE is minimized when $p(X) = E[A|X]$.

**The Fairness Criterion.** However, in insurance policy pricing, the precision criterion is not the sole part of the picture; just as important is that the estimated premia do not systematically discriminate against specific segments of the population. We call this objective the fairness criterion. We define the bias of the premia $b(P)$ to be the difference between the average premium and the average incurred amount, in a given population $P$:

$$b(P) = \frac{1}{|P|} \sum_{(x_i, a_i) \in P} p(x_i) - a_i, \quad (3)$$

where $|P|$ denotes the cardinality of the set $P$, and $p(\cdot)$ is some premia estimation function. A possible fairness criterion would be based on minimizing the norm of the bias over every subpopulation $Q$ of $P$. From a practical standpoint, such a minimization would be extremely difficult to carry out. Furthermore, the bias over small subpopulations is hard to estimate with statistical significance. We settle instead for an approximation that gives good empirical results. After training a model to minimize the MSE criterion (2), we define a finite number of disjoint subsets (subpopulations) of the test set $P$, $P_k \subset P$, $P_k \cap P_{k'} = \emptyset$, and verify that the absolute bias is not significantly different from zero. The subsets $P_k$ can be chosen at convenience; in our experiments, we considered 10 subsets of equal size delimited by the deciles of the test set premium distribution. In this way, we verify that, for example, for the group of contracts with a premium between the 5th and the 6th decile, the average premium matches the average claim amount.

3 Models Evaluated

An important requirement for any model of insurance premia is that it should produce positive premia: the company does not want to charge negative money to its customers! To obtain positive outputs neural networks we have considered using an exponential activation function at the output layer but this created numerical difficulties (when the argument of the exponential is large, the gradient is huge). Instead, we have successfully used the “soft plus” activation function (Dugas et al., 2001):

$$\text{soft plus}(s) = \log(1 + e^s)$$

where $s$ is the weighted sum of an output neuron, and soft plus($s$) is the corresponding predicted premium. Note that this function is convex, monotone increasing, and can be considered as a smooth version of the “positive part” function $\max(0, x)$.

The best model that we obtained is a mixture of experts in which the experts are positive outputs neural networks. The gates network (Jacobs et al., 1991) has softmax outputs to obtain positive weights summing to one.