Bayesian Factorization Machines for Risk Management and Robust Decision Making

Angulo, P., Gallego, V., Gómez-Ullate, D. and Suárez-García, P.

Abstract When considering different allocations of the marketing budget of a firm, some predictions, that correspond to scenarios similar to others observed in the past, can be made with more confidence than others, that correspond to more innovative strategies. Selecting a few relevant features of the predicted probability distribution leads to a *multi-objective optimization* problem, and the Pareto front contains the most interesting media plans. Using expected return and standard deviation we get the familiar two moment decision model, but other problem specific additional objectives can be incorporated. *Factorization Machines*, initially introduced for recommendation systems, but later used also for regression, are a good choice for incorporating interaction terms into the model, since they can effectively exploit the sparse nature of typical datasets found in econometrics.

1 Introduction

In the marketing industry, a batch of advertising slots is bought on a yearly basis, as this allows better pricing from the media retailer. Our problem is to assist a one-time decision that involves distributing a fixed advertising budget over a one year period. On each week, advertising budget can go into several *channels*, such as TV, Radio or Out-of-Home. A *strategy* is a choice of investment on each week and channel.

Angulo, Pablo
ETSIN, UPM, Avd. de la Memoria 4, Madrid, e-mail: pablo.angulo@upm.es

Gallego, Víctor
ICMAT, C/ Nicolás Cabrera 13-15, Campus de Cantoblanco e-mail: victor.gallego@icmat.es

Gómez-Ullate, David
ICMAT, C/ Nicolás Cabrera 13-15, Campus de Cantoblanco, e-mail: david.gomez-ullate@icmat.es

Suárez-García, Pablo
Depto. Física Teórica, Facultad de Física, UCM e-mail: pasuarez@fis.ucm.es

The decision should be based on historic data: a time series of N observations $y_t \in \mathbb{R}^R$ for time periods $t = 1,...,N$, plus, for each time period, a set of predictor variables x_t^i that media specialists believe to be correlated with sales. For one particular customer, a chain of fast food restaurants, y_t^r are the total sales of restaurant $r = 1 \dots R$ in week $t = 1 \dots N$, and the predictor variables represent the effect of climate, sport events, special holidays, socioeconomic indicators such as unemployment or inflation, and of course the investments in advertisements during that week. We split the predictor variables into two sets:

- the variables that we cannot control: weather, events, economic indicators, etc... Some of them are real valued (unemployment, mean temperature, ...), while others are binary (Christmas, Easter, major sport event, ...). The holiday type and events variables are *sparse*, since most of them are zero at any given week.
- the variables that we control, i.e. the variables that specify an investment strategy. All of them are real and positive.

We know the values of some of the predictor variables in the first set with certainty (events and holidays variables), while for others we only have probabilistic forecasts (weather and socioeconomic variables). We are allowed to fix the investment strategy, given some constraints such as total budget.

In [2], we predict sales for the next week, as a function of investmentes, with the knowledge of all previous weeks, and adapt the control variables each week. This setup is now being used at media analytics company Annalect, who ordered this study.

2 Prediction

We are given a set of N observations, $\{(x_1,y_1),...,(x_N,y_N)\}$, where $x_t \in X$ is the feature vector of the t-th week and $y_t \in \mathbb{R}^R$ is the target: the sales y_t^j at each of the $j = 1 \dots R$ restaurants and each time $t = N+1 \dots N+t$.

Factorization Machines [6] use a quadratic function where the matrix for the quadratic part has rank at most k:

$$g(x) := w_0 + \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j=i+1}^{p} \langle v_i, v_j \rangle x_i x_j \tag{1}$$

- w_0 is the global bias.
- w_i captures the effect of the $i-$th feature.
- $\langle v_i, v_j \rangle$ captures the interaction effect between features i and j, but using one latent factor $v_i \in \mathbb{R}^k$ per feature.

FMs requires $\sim kp$ parameters, while the full-rank second order version needs $\sim p^2/2$. This is a critical aspect of FMs making it suitable for fitting small size datasets that arise in many business contexts. Factorization machines in particular, and factorization models in general, have been widely employed in tasks such as

recommender systems or ad click prediction [6, 1, 4], problems characterised by the prevalence of outrageously big datasets. We show that these models may also be helpful for other kind of datasets in which observations are scarce and there are sparse blocks in the data matrix X.

We add another block of predictor variables: x^r are binary, and there is one for each restaurant $r = 1 \dots R$, so that exactly one of them takes the value 1 at any data point. The model identifies each restaurant with its mean w_i and its feature vector $v_i \in \mathbb{R}^k$, and this forces the model to generalize.

In *Bayesian Parametric Regression*, the mean of the target variable is a deterministic function of the predictor variables, but the function depends on a few unknown parameters. We assume that the function belongs to the FM family and the distribution is a normal with fixed variance (that we estimate later).

$$y = w_0 + \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j=i}^{p} \langle v_i, v_j \rangle x_i x_j + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

In other words, the likelihood of y conditioned to the model parameters is

$$p(y|w_0, w_i, v_i) \sim \exp\left(\left(\frac{y - w_0 + \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j=i}^{p} \langle v_i, v_j \rangle x_i x_j}{\sigma}\right)^2\right) \tag{3}$$

Since the parameters are unknown, we model our uncertainty about them with a prior probability distribution, and Bayes theorem gives the posterior belief about the model parameters. For fixed values of the predictor variables and the model parameters, (2) gives the sales of the restaurants. We integrate our posterior probability measure over the set of parameters and we also integrate over the distributions of the predictor variables that we don't known with certainty. If we add the sales of all the restaurants, we get a probability distribution for a single real number.

A straightforward application of Bayes theorem leads to untractable integrals, so we use the *Markov Chain Monte Carlo* (MCMC) method [3, ch 12]. MCMC replaces the probability measure by a representative sample that is obtained by performing a random walk on the feature space X, but one that is modulated by a multiple of the posterior density, which can be computed as the product of the prior distribution and the likelihood.

The optimal decision problem relies heavily on our ability to make good forecasts of the sales in the future, not only of its expectation but also of the variance and other features of its probability distribution. With the aim of measuring the quality of our predictions, we follow the standard procedure of spliting the data set into a training and a test set. The first one is used to learn the posterior probability distribution whereas the test set is employed to estimate the performance of the model.

We can compare the posterior mean with the observed sales to get a first measure of the quality of the predictions, but we must also calibrate our estimations of the variance. For any week and restaurant in the test set, our prediction is a different probability distribution, and we only get one sales value for each such distribution.

In order to solve this, we apply the probability integral transform, that takes any probability measure into the uniform distribution in the interval $[0,1]$. We get in this way a sample that we can compare to the $[0,1]$-uniform distribution, both for measuring goodness-of-fit and for selecting hyperparameters.

3 Multiobjective Optimization

With this predictive model we consider a multiobjective optimization problem over the control variables. The typical choice for multiobjective optimization function in financial settings is maximizing expected sales while minimizing expected variance, but other problem specific additional objectives can be incorporated.

For the chain of fast food restaurants, we added the *"restaurants at risk"* metric (RAR). An innovative strategy might increase total expected sales by increasing sales in a few big restaurants, but at the same time dissapoint many restaurant owners, who believe that the chosen strategy harms their restaurant in particular.

The RAR is the expected number of restaurants whose sales figure with the new plan will be below the 5% quantile of the base plan. The RAR is not zero (but 5%!) if the new plan is actually the same as the base plan.

In order to find the Pareto frontier, we use the technique of *scalarization*, in which the different objectives are combined into a single function in different ways. There are many alternative methods [5], but the simplest weighted sum method was good enough: maximize a linear function of the multiple objectives with different weights, and vary the weights to get new points in the Pareto frontier.

In the end, the outcome of our model is a representative set of Pareto optimal investment strategies for the set of objective functions. The human decision maker can choose among this small set of concrete strategies, which is more convenient than ellicitation of the full utility function.

References

1. Freudenthaler C, Schmidt-Thieme L, Rendle S (2011) Bayesian Factorization Machines. In: Workshop on Sparse Representation and Low-rank Approximation, Neural Information Processing Systems (NIPS), Granada, Spain.
2. Gallego V, Angulo P, Suárez-Garcia P, Gómez-Ullate D (2018) . Sales forecasting and risk management under uncertainty in the media industry. http://arxiv.org/abs/1801.03050
3. Gelman A, Carlin J B, Stern H S, Dunson D B, Vehtari A, and Rubin D B (2014) Bayesian Data Analysis, Third Edition. CRC Press.
4. Juan Y, Zhuang Y, Chin W S, Lin C J (2016) Field-aware Factorization Machines for CTR Prediction. In: ACM (ed) Procs. of the 10th ACM Conf. on Recommender Systems, 43–50.
5. Marler R T, Arora J S (2004) Survey of multi-objective optimization methods for engineering In: Struct Multidisc Optim, doi: 10.1007/s00158-003-0368-6
6. Rendle S (2010) Factorization Machines. In: IEEE Computer Society (ed) Proceedings of the 2010 IEEE International Conference on Data Mining, 995–1000.