Expert Knowledge Elicitation for *large p, small n* Regression

Marta Soare, Muhammad Ammad-ud-din, Samuel Kaski
Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University

Context. We consider regression under the "extremely small n large p" condition. In particular, we focus on problems with so small sample sizes n compared to the dimensionality p, that predictors cannot be estimated without prior knowledge. This setup occurs in personalized medicine, for instance, when predicting treatment outcomes for an individual patient based on noisy high-dimensional genomics data. The few observations and the information on the structure in the data are in this case not sufficient. A remaining source of information is knowledge of experts which can have a major impact when extracted and used efficiently. For concreteness, in the sequel we describe the problem formulation with terminology of treatment effectiveness prediction but the setup is naturally more generally applicable.

Goal. The goal is to improve prediction of the effect of a treatment on a target patient, by using feedback provided by an expert. Assume a small set of observation data, which can be used to learn an initial predictor. The set consists of n observed treatment responses coming from patients who had previously received the same treatment. Denote the matrix of genomic features with $\mathcal{X}$, where the size of $\mathcal{X}$ is $n \times p$ ($p \gg n$), and on each row $i = 1, \ldots, n$ we have the p genomic features corresponding to patient x_i. For a "target patient" x^* the same genomic measurements are available and the goal is to predict as accurately as possible the corresponding treatment response.

Linear Regression. In a first model we assume there is a linear relation between the medical features of the patient and the expected result of the treatment. More precisely, the observed response for each patient i is given by $y_i = x_i \theta^{*\top} + \eta$, where $\theta^* \in \mathbb{R}^p$ is an unknown parameter underlying the linear function and η is i.i.d noise, quantifying the inherent noise in the measurements of the drug effects for each patient. The coordinate $\theta^*(j)$ of the parameter vector encodes the weight or relevance that each feature j has in computing the treatment response. We define the loss as the expected quadratic loss for the target patient.

Expert knowledge. We assume that the expert is able to report unbiased feedbacks on the values of a subset of the features of $\theta^*(i)$. This assumption is very simplifying in the personalized medicine case, and requires that the expert either has important additional knowledge of the particular patient, or is able to use his/her expertise to infer the correct value from the shown data x^* and the initial weight vector estimated based on the other patients. We assume this type of expert knowledge to be expensive and we hence place a strict restriction on the number of features m on which the expert can provide feedback. Thus, the research problem we address is to identify the $m << p$ most informative features on which to elicit expert knowledge such that the improvement in the treatment effect prediction is maximized.

Preliminary results. We propose to learn the regression parameters in two stages. First, an initial estimate $\hat{\theta}_{\text{init}}$ is learnt on the "large p, small n" training data with appropriate regularization, efficiently capturing the information in that data set. The estimate is then improved by using an elicitation strategy which identifies and asks expert feedback on the most informative features. We derive conditions under which the elicitation strategy is optimal for predicting the response of a target patient. Using simulated experts and the data from the GDSC genomics dataset, we compare the performance of our algorithm to that of a baseline strategy which does not use expert knowledge and estimates the drug response only based on $\hat{\theta}_{\text{init}}$. The improvement obtained when using expert knowledge elicitation grows with the number of observed feedbacks. The results hold for noisy experts and when expert knowledge is restricted to a subset of features.

Future work. We have introduced a novel setup that brings together expert elicitation and the difficult "large p, small n" regression problem. Starting from noisy estimates based on extremely small sample sizes, we demonstrated the prediction improvement that can be obtained by efficiently using few expert feedbacks. This simplified problem setting is intended to be a starting point that opens up new interesting theoretical questions and a line of applied work towards new solutions for the "small n, large p" settings, particularly useful in the very timely problem of personalized medicine.