## Abstract

*Ongoing title:*
***Interactive user modelling for human-in-the-loop machine learning***

## Abstract

*En puhu suomea.*

# Preface

To fill.

Espoo, August 5, 2019,

Pedram Daee

# Contents

# List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

**I** Pedram Daee, Joel Pyykkö, Dorota Głowacka, and Samuel Kaski. Interactive Intent Modeling from Multiple Feedback Domains. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, Sonoma, California, USA, 71–75, March 2016.

**II** Pedram Daee*, Tomi Peltola*, Marta Soare*, and Samuel Kaski. Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. *Machine Learning*, 106, 9-10, 1599–1620, 2017.

**III** Iiris Sundin*, Tomi Peltola*, Luana Micallef, Homayun Afrabandpey, Marta Soare, Muntasir Mamun Majumder, Pedram Daee, Chen He, Baris Serim, Aki Havulinna, Caroline Heckman, Giulio Jacucci, Pekka Marttinen, and Samuel Kaski. Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge. *Bioinformatics*, 34, 13, i395–i403, 2018.

**IV** Pedram Daee*, Tomi Peltola*, Aki Vehtari, and Samuel Kaski. User Modelling for Avoiding Overfitting in Interactive Knowledge Elicitation for Prediction. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, Tokyo, Japan, 305–310, March 2018.

**V** Giulio Jacucci, Oswald Barral, Pedram Daee, Markus Wenzel, Baris Serim, Tuukka Ruotsalo, Patrik Pluchino, Jonathan Freeman, Luciano Gamberini, Samuel Kaski, Benjamin Blankertz. Integrating Neurophysiological Relevance Feedback in Intent Modeling for Information Retrieval. *Journal of the Association for Information Science and Technology*, 2019.

# Author's Contribution

### Publication I: "Interactive Intent Modeling from Multiple Feedback Domains"

The author had the main responsiblity in problem formulation and modeling. The author designed and implemented the simulation experiment. Joel Pyykkö and the author built the system for user studies and conducted them together. The author wrote the initial draft of the manuscript, after which all co-authors joined for revisions.

### Publication II: "Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction"

The ideas and experiments in this article were designed jointly (the first three authors contributed equally). The author had the main responsibility in the derivation of the sequential experimental design and implementation of the experiments. Dr. Tomi Peltola derived and implemented the posterior approximation. The manuscript was written jointly.

### Publication III: "Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge"

The author contributed on formulation of the sequential experimental design and implementation of a portion of the early version of the experiments. The author made comments to the manuscript in preparation.

### Publication IV: "User Modelling for Avoiding Overfitting in Interactive Knowledge Elicitation for Prediction"

The ideas and experiments in this article were designed jointly (the first two authors contributed equally). The author designed and implemented the user study. Dr. Tomi Peltola had the main responsibility of the model formulation. The first two authors wrote the initial draft of the manuscript, after which all co-authors joined for revisions.

### Publication V: "Integrating Neurophysiological Relevance Feedback in Intent Modeling for Information Retrieval"

The author had the main responsibility in design and implementation of the interactive intent modelling and information retrieval system, and writing of the correspnding sections. All the authors contributed to paper revisions.

# 1. Introduction

Whether it is an everyday user searching for an application in her mobile phone or a doctor working with a cancer diagnostic system, humans and machines are increasingly interacting with each other. The goal of the thesis is to improve this interaction by incorporating a model of the human user in the system they are interacting with. In particular, the thesis considers the family of problems where the human and machine interact to solve a prediction problem. Such problems can include personalized search activity or medical prediction about the response of a cancer drug. An important common factor in both these scenarios is that the number of labeled data (training data) that the machine can use to make predictions, is usually very few compared to the dimension of search space. This results in ill-posed statistical learning since there are limits in how low in sample size statistical methods can go [1].

## 1.1 Motivation

## 1.2 Research questions and contributions

This thesis investigates methods to tackle the limited user interaction challenge in interactive machine learning for prediction. The thesis focuses on scenarios where there is few labeled data available compared to the dimension of the problem, or when a human user is provider of the labeled data. The core idea of the thesis is to jointly model the human user with the data as part of a unified probabilistic model and use the model to improve the interaction.

**RQ1 –** *Can we exploit new sources of interaction as additional learning signals from human user to improve interactive intent modelling?*

Publications I and V contribute to this research question by proposing models to incorporate new types of user feedback to amend the limited feedback in exploratory information seeking tasks. The tasks considered are document

search scenarios where a user needs to sequentially provide relevance feedback to suggested keywords in order to find the targeted document. This is modelled as a multi-armed bandit problem with the goal of finding the most relevant document with minimum interaction. In particular, Publication I couples user relevance feedback on both documents and keywords by assuming a shared underlying latent model connected through a probabilistic model of the relationship between keywords and documents. Thompson sampling on the posterior of the latent intent was then used to recommend new documents and keywords in each iteration. Publication V investigates the use of implicit relevance feedback from neurophysiology signals for effortless information seeking. The work contributes by demonstrating how to integrate this inherently noisy and implicit feedback source with scarce explicit interaction. A model for controlling the accuracy of the feedback given its nature (implicit or explicit) was introduced. Similar to Publication I, Thompson sampling was used to control the exploration and exploitation balance of the recommendations. Both publications were evaluated by user studies in realistic information seeking tasks.

**RQ2 –** *Can expert knowledge about high dimensional data models be elicited to improve the prediction performance?*

Publications II and III contribute to this research question. Publication II proposes a framework for user knowledge elicitation as a probabilistic inference process, where the user knowledge is sequentially queried to improve predictions. In particular, sparse linear regression is considered as the data model with access to only few high-dimensional training data. It is assumed that there are experts who have knowledge about the relevance of the covariates, or of values of the regression coefficients and can provide this information to the data model if queried. The work contributes by an algorithm and computational approximation for fast and efficient interaction, which sequentially identifies the most informative queries to ask from the user. Publication III, builds on Publication II by adding user knowledge about direction of relevance of covariates and applying the method in important applications of precision medicine with the goal of predicting the effects of different treatments using high-dimensional genomic measurements. Both publications were evaluated by extensive simulations and user studies. Source codes for methods presented in Publications II and III, and user study data from Publication II are available at `https://github.com/HIIT/knowledge-elicitation-for-linear-regression` and `https://github.com/AaltoPML/knowledge-elicitation-for-precision-medicine`.

**RQ3 –** *Is it enough to incorporate human knowledge directly in the data model as explained in RQ2, or could it be beneficial to account for rational knowledge updates that humans may undergo during the interaction?*

Publication IV contributes to this research question by modelling the knowledge provider, here the human user, as a rational agent that updates its knowledge about the underlying prediction task during the interaction. In particular, certain aspects of training data may be revealed to the user during knowledge elicitation. The design of the system is then critical, since the elicited

user knowledge cannot be assumed to be independent from the data model knowledge coming from the training data. If not accounted properly, knowledge elicitation can lead to double use of data and overfitting, if the user reinforces noisy patterns in the data. We propose a user modelling methodology, by assuming simple rational behaviour, to correct the problem and evaluate the method in a user study. Source code and user study data are available at `https://github.com/HIIT/human-overfitting-in-IML`.

## 1.3 Organization of the thesis

The organization of the thesis is as follows. Chapter 2 provides an overview of probabilistic modelling and introduces our approach of modelling the user and data as a joint probabilistic model. Chapter three investigates the purpose of interactions and reviews different utility functions designed to minimize the effort of the user. The fourth chapter summarizes Publications I-IV. Chapter five concludes the thesis and provides discussions for future works.

# 2. Probabilistic modelling of data and user

This chapter provides a brief introduction to probabilistic modelling as the main statistical framework that is used through the thesis. After some preliminaries, Section 2.2 briefly reviews the type of linear models that are used in Publication I-IV for prediction. Section 2.3 introduces different types of user interaction with the linear model and explains how user knowledge about the model can be incorporated as observational feedback. Finally, the key components for modelling observational data and user feedback as a joint probabilistic model is introduced in Section 2.4.

## 2.1 Preliminaries

The core idea of probabilistic modelling is to describe all the unobserved parameters and observed data as random variables from probability distributions. The unobserved parameters include the unknown quantity of interest or other parameters that affect the data or the quantity of interest. Bayesian inference provides a powerful framework to fit the described probabilistic model to observational data [2]. A core feature of Bayesian inference is that it provides probability distributions as the solution, compared to deterministic methods which provide a single outcome. This uncertainty quantification is of particularly high interest in cases where few observational data are available or when the data acquisition scheme is controlled by the model. Both of these constrains are prominently present in the tasks investigated by this thesis.

We follow the notation of [2] and use $p(.)$ to denote a probability distribution and $p(.|.)$ a conditional distribution. Consider the case where there are a set of observations $\mathscr{D} = \{y_1, \ldots, y_n\}$ generated from a probabilistic model with an unobserved parameter of interest $\theta$. The observational model for an observation $y$ describes the conditional density of $y$ given the parameter $\theta$ and is denoted as $y \sim p(y \mid \theta)$. It is usually assumed that the observations are conditionally independent given $\theta$, enabling us to write the model for all observations as $p(\mathscr{D} \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta)$. The observational model is called likelihood function if perceived as a function of $\theta$ with fixed observation $y$. One of the core questions

in statistical inference is to estimate the parameter of interest $\theta$ based on observations $\mathscr{D}$. Bayesian inference answers this question by computing the conditional distribution of $\theta$ given $\mathscr{D}$ following the Bayes rule

$$p(\theta \mid \mathscr{D}) = \frac{p(\mathscr{D} \mid \theta)p(\theta)}{p(\mathscr{D})}. \tag{2.1}$$

Where $p(\theta)$ represents the prior belief about $\theta$ and $p(\mathscr{D})$ is called the marginal likelihood and acts as a normalization factor as it does not depend on $\theta$. The marginal likelihood can be computed using the marginalization rule, i.e., $p(\mathscr{D}) = \int_\theta p(\mathscr{D},\theta)d\theta = \int_\theta p(\mathscr{D} \mid \theta)p(\theta)d\theta$ [1]. $p(\theta \mid \mathscr{D})$ is called the posterior and it expresses the uncertainty surrounding the true value of $\theta$, after updating the prior assumptions about $\theta$ (i.e., the prior distribution) with the knowledge coming from the observations through the likelihood.

In many cases, we may be more interested to make a prediction about an unknown observable data point $\hat{y}$ rather than the parameter $\theta$. Bayesian inference allows us to compute the conditional distribution of this unknown observable data given the observed data point as

$$p(\hat{y} \mid \mathscr{D}) = \int_\theta p(\hat{y} \mid \theta)p(\theta \mid \mathscr{D})d\theta. \tag{2.2}$$

$p(\hat{y} \mid \mathscr{D})$ is known as the posterior predictive distribution and represents the uncertainty about a potential new observation. An example usage of this distribution could be when we want to predict the value of a test data. Since test data is not observed, we can use posterior predictive distribution as our best guess. However, in many applications, only a value (and not a distribution) is required as the prediction. This can be handled by using some statistics of the distribution (for example mean or median) as the prediction. Still, the quantified uncertainty in the distribution can be useful as it provides knowledge about how certain we are about our estimate. Particularly, this uncertainty can help us to design more efficient interaction with a user, as will be discussed in Section 3.

## 2.2 Modelling data for prediction

Prediction is one of the core problems in statistical analysis and supervised machine learning. Given a set of $n$ input and output pairs, called training data, denoted by $\mathscr{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, the goal is to find a mapping from inputs to outputs. Here, $\boldsymbol{x}_i = [x_i^1, \ldots, x_i^d]^\top \in \mathbb{R}^d$ is a d-dimensional column vector representing the values of $i^{th}$ data point [2]. The dimensions are commonly called feature, covariates, or attribute. $y_i$ is the corresponding response (or target) variable which can be anything depending on the underlying problem. In this thesis, we consider the regression tasks, meaning that we model response variables by real values,

[1] The integral turns to summation for discrete $\theta$.

[2] We use subscripts to index an specific item (e.g, one particular observation out of several) and superscripts to refer to dimensions of the variable.

i.e., $y_i \in \mathbb{R}$. The problem is called classification in supervised learning if the response variable is restricted to categorical values (sometimes called classes).

A well-studied and widely practical type of regression, known as linear regression, assumes that the relationship between all inputs and their corresponding response variables is linear. This relationship can be written as

$$y_i = \sum_{j=1}^{d} x_i^j w^j + \epsilon_i = \boldsymbol{x}_i^\top \boldsymbol{w} + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $\boldsymbol{w} \in \mathbb{R}^d$ is the regression coefficients or the model's weights and $\epsilon_i$ is the residual error between linear prediction $\boldsymbol{x}_i^\top \boldsymbol{w}$ and the response value $y_i$. Given the labeled data $\mathscr{D}$, a commonly used error function for this problem is the sum of squared of residual errors $\sum_{i=1}^{n} (\boldsymbol{x}_i^\top \boldsymbol{w} - y_i)^2$. By stacking the inputs in $\boldsymbol{X} = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times d}$ and outputs in $\boldsymbol{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$ the error can be written as $(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^\top (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$. The frequentist solution directly finds the point estimate for $\boldsymbol{w}$ that minimizes this error. It is straightforward to show that $\hat{\boldsymbol{w}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ would be the point estimate solution given that $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ is invertible.

However, as mentioned, we are interested in directly quantifying the uncertainty of the solution. The probabilistic way to model this problem is to explicitly describe the model assumptions (likelihood and priors) as probability distributions. In linear regression, the residual errors $\epsilon_i$ and the weights $\boldsymbol{w}$ are modelled as random variables and the inputs $\boldsymbol{x}_i$ as vector of values that are given. A customary assumption is to model the residual errors as independent zero-mean Gaussian random variables $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, where $\sigma^2$ is variance of the Gaussian distribution which indicates the model tolerance about residual errors. It is common to also model $\sigma^2$ as another random variable with its own distribution assumption, however, we can consider it to be a fixed hyperparameter for simplicity for now. The quantity of interest in the linear model is the regression coefficients. To complete the Bayesian inference loop, we need to consider a prior distribution on $\boldsymbol{w}$. There are many ways to do this depending on the underlying task. One simple prior could be to assume that coefficients are independent and each come from a zero-mean Gaussian distribution, encouraging the weights to be close to zero. This simple Bayesian linear regression can be described by

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{w}, \sigma^2 \mathbf{I})$$

$$\boldsymbol{w} \sim \mathrm{N}(0, \tau^2 \mathbf{I}),$$

where $\tau^2$ is the variance of the weights and $\mathbf{I}$ is the identity matrix. For simplicity we assume for now that $\tau^2$ is also a fixed hyperparameter. Therefore, The only unobserved parameters of this model is $\boldsymbol{w}$. The Bayesian rule (Equation 2.1) allows us to derive the posterior for $\boldsymbol{w}$ as

$$p(\boldsymbol{w} \mid \mathscr{D}) = \frac{p(\mathscr{D} \mid \boldsymbol{w}) p(\boldsymbol{w})}{p(\mathscr{D})}. \tag{2.3}$$

Generally, the posterior in many problems cannot be analytically derived (we will discuss this issue more in XXX). For this simple model, however, an analytical solution is available. We will go through the steps as a hand on with posterior inference. Writing down the distribution assumptions and dropping $p(\mathscr{D})$ (as it is a constant factor) gives

$$p(\boldsymbol{w} \mid \mathscr{D}) \propto \prod_{i=1}^{n} \mathrm{N}(y_i | \boldsymbol{x}_i^\top \boldsymbol{w}, \sigma^2) \mathrm{N}(\boldsymbol{w} | 0, \tau^2 \mathbf{I}). \tag{2.4}$$

Note that we use for example $\boldsymbol{w} \sim \mathrm{N}(0, \tau^2 \mathbf{I})$ to denote the random variable $\boldsymbol{w}$ and $\mathrm{N}(\boldsymbol{w}|0, \tau^2 \mathbf{I})$ to refer to the density function.

HERE GOES THE REST OF STEPS

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{w}, \sigma^2 \mathbf{I}), \tag{2.5}$$

ALSO EXPLAIN BRIEFLY THE NON BAYESIAN TREATMENT AND EXPLAIN WHY WE GO WITH BAYES/

### 2.2.1 Bayesian Linear regression

### 2.2.2 Sparse priors

## 2.3 Modelling the user

## 2.4 Key components of the joint model

Let $y$ and $x$ denote the outputs (target variables) and inputs (covariates), and $\theta$ and $\phi_y$ the model parameters. Let $f$ encode input (*feedback*) from the user, presumably a domain expert, and $\phi_f$ be model parameters related to the user input. We identify the following key components:

1. An observation model $p(y \mid x, \theta, \phi_y)$ for $y$.

2. A feedback model $p(f \mid \theta, \phi_f)$ for the expert's knowledge.

3. A prior model $p(\theta, \phi_y, \phi_f)$ completing the hierarchical model description.

4. A query algorithm and user interface that facilitate gathering $f$ iteratively from the expert.

5. Update process of the model after user interaction.

The observation model can be any appropriate probability model. It is assumed that there is some parameter $\theta$, possibly high-dimensional, that the expert has knowledge about. The expert's knowledge is encoded as (possibly partial) feedback $f$ that is transformed into information about $\theta$ via the feedback model. Of course, there could be a more complex hierarchy tying the observation and feedback models, and the feedback model can also be used to model more user-centric issues, such as the quality of or uncertainty in the knowledge or user's interests.

# 3. User interaction with the probabilistic model

## 3.1 Active learning and experimental design

## 3.2 Multi-armed bandits and Bayesian optimization

# 4. Summary of the Contributions

This chapter briefly summarizes the contributions of Publications I-V with emphasize on answering the research questions of the thesis.

## 4.1 Interactive intent modelling from multiple feedback domains (Publications I and V)

## 4.2 Expert knowledge elicitation for high-dimensional prediction (Publications II and III)

## 4.3 User modelling for avoiding overfitting in knowledge elicitation (Publication IV)

# 5. Discussion

# References

[1] David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

[2] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman & Hall/CRC, 3rd edition, 2014.

# Publication I

Pedram Daee, Joel Pyykkö, Dorota Głowacka, and Samuel Kaski. Interactive Intent Modeling from Multiple Feedback Domains. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, Sonoma, California, USA, 71–75, March 2016.

# Publication II

Pedram Daee*, Tomi Peltola*, Marta Soare*, and Samuel Kaski. Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. *Machine Learning*, 106, 9-10, 1599–1620, 2017.

# Publication III

Iiris Sundin*, Tomi Peltola*, Luana Micallef, Homayun Afrabandpey, Marta Soare, Muntasir Mamun Majumder, Pedram Daee, Chen He, Baris Serim, Aki Havulinna, Caroline Heckman, Giulio Jacucci, Pekka Marttinen, and Samuel Kaski. Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge. *Bioinformatics*, 34, 13, i395–i403, 2018.

# Publication IV

Pedram Daee*, Tomi Peltola*, Aki Vehtari, and Samuel Kaski. User Modelling for Avoiding Overfitting in Interactive Knowledge Elicitation for Prediction. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, Tokyo, Japan, 305–310, March 2018.

# Publication V

Giulio Jacucci, Oswald Barral, Pedram Daee, Markus Wenzel, Baris Serim, Tuukka Ruotsalo, Patrik Pluchino, Jonathan Freeman, Luciano Gamberini, Samuel Kaski, Benjamin Blankertz. Integrating Neurophysiological Relevance Feedback in Intent Modeling for Information Retrieval. *Journal of the Association for Information Science and Technology*, 2019.