

Abstract

Ongoing title:

Interactive user modelling for human-in-the-loop machine learning

Abstract

En puhu suomea.

Preface

To fill.

Espoo, August 7, 2019,

Pedram Daee

Contents

Preface	3
Contents	5
List of Publications	7
Author's Contribution	9
1. Introduction	11
1.1 Motivation	11
1.2 Research questions and contributions	11
1.3 Organization of the thesis	13
2. Probabilistic modelling of data and user	15
2.1 Preliminaries	15
2.2 Modelling data for prediction	16
2.2.1 Bayesian Linear regression	17
2.3 Modelling the user	20
2.4 Key components of the joint model	20
3. User interaction with the probabilistic model	21
3.1 Active learning and experimental design	21
3.2 Multi-armed bandits and Bayesian optimization	21
4. Summary of the Contributions	23
4.1 Interactive intent modelling from multiple feedback domains (Publications I and V)	23
4.2 Expert knowledge elicitation for high-dimensional prediction (Publications II and III)	23
4.3 User modelling for avoiding overfitting in knowledge elicitation (Publication IV)	23
5. Discussion	25

Contents

References **27**

Publications **29**

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 Daei, Joel Pyykkö, Dorota Glowacka, 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 Daei^{*}, 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** Iris Sundin^{*}, Tomi Peltola^{*}, Luana Micallef, Homayun Afrabandpey, Marta Soare, Muntasir Mamun Majumder, Pedram Daei, 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 Daei^{*}, 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 Daei, 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 responsibility 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 corresponding 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 constraints are prominently present in the tasks investigated by this thesis.

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

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

$$p(\theta | \mathcal{D}) = \frac{p(\mathcal{D} | \theta)p(\theta)}{p(\mathcal{D})}. \quad (2.1)$$

Where $p(\theta)$ represents the prior belief about θ and $p(\mathcal{D})$ is called the marginal likelihood and acts as a normalization factor as it does not depend on θ . The marginal likelihood can be computed using the marginalization rule, i.e., $p(\mathcal{D}) = \int_{\theta} p(\mathcal{D}, \theta) d\theta = \int_{\theta} p(\mathcal{D} | \theta) p(\theta) d\theta$ ¹. $p(\theta | \mathcal{D})$ is called the posterior and it expresses the uncertainty surrounding the true value of θ , after updating the prior assumptions about θ (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 θ . Bayesian inference allows us to compute the conditional distribution of this unknown observable data given the observed data point as

$$p(\hat{y} | \mathcal{D}) = \int_{\theta} p(\hat{y} | \theta) p(\theta | \mathcal{D}) d\theta. \quad (2.2)$$

$p(\hat{y} | \mathcal{D})$ 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 $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the goal is to find a mapping from inputs to outputs. Here, $\mathbf{x}_i = [x_i^1, \dots, x_i^d]^T \in \mathbb{R}^d$ is a d -dimensional column vector representing the values of i^{th} data point². 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,

¹The integral turns to summation for discrete θ .

²We 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 (called classes).

2.2.1 Bayesian Linear regression

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 described as

$$y_i = \sum_{j=1}^d x_i^j w^j + \epsilon_i = \mathbf{x}_i^\top \mathbf{w} + \epsilon_i, \quad i = 1, \dots, n,$$

where $\mathbf{w} \in \mathbb{R}^d$ is the regression coefficients or the model's weights and ϵ_i is the residual error between linear prediction $\mathbf{x}_i^\top \mathbf{w}$ and the response value y_i . Given the labeled data \mathcal{D} , a commonly used error function is the sum of squared of residual errors $\sum_{i=1}^n (\mathbf{x}_i^\top \mathbf{w} - y_i)^2$. By stacking the inputs in $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times d}$ and outputs in $\mathbf{y} = [y_1, \dots, y_n]^\top \in \mathbb{R}^n$ the error can be expressed in vector format as $(\mathbf{X}\mathbf{w} - \mathbf{y})^\top (\mathbf{X}\mathbf{w} - \mathbf{y})$. The frequentist solution directly finds the point estimate for \mathbf{w} that minimizes this error (also known as least squares solution). It is straightforward to show that $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ would be the point estimate solution given that $(\mathbf{X}^\top \mathbf{X})^{-1}$ 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 ϵ_i and the weights \mathbf{w} are modelled as random variables and the inputs \mathbf{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 $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is variance of the Gaussian distribution which indicates the model tolerance about residual errors. It is common to also model σ^2 as another random variable with its own distribution assumption, however, we consider it to be a fixed hyperparameter for simplicity. 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 \mathbf{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

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

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

where τ^2 is the variance of the weights and \mathbf{I} is the identity matrix. For simplicity we assume that τ^2 is also a fixed hyperparameter. Therefore, The only

unobserved parameters of this model is \mathbf{w} . The Bayesian rule allows us to derive the posterior for \mathbf{w} as³

$$p(\mathbf{w} | \mathcal{D}) = \frac{p(\mathcal{D} | \mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} = \frac{N(\mathbf{y} | \mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I})N(\mathbf{w} | 0, \tau^2 \mathbf{I})}{p(\mathcal{D})}. \quad (2.3)$$

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 exercise with posterior inference. Before starting, it would be useful to note that a multivariate Gaussian distribution can be expressed as

$$\begin{aligned} N(\mathbf{w} | \mu, \Sigma) &\propto \exp(\mathbf{w} - \mu)^\top \Sigma^{-1} (\mathbf{w} - \mu) \\ &\propto \exp(\mathbf{w}^\top \Sigma^{-1} \mathbf{w} - 2\mathbf{w}^\top \Sigma^{-1} \mu), \end{aligned} \quad (2.4)$$

after dropping all the terms that are constant with respect to \mathbf{w} .

To derive the posterior, we follow Equation 2.3

$$\begin{aligned} p(\mathbf{w} | \mathcal{D}) &\propto \exp\left(\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w})\right) \exp\left(\frac{1}{2\tau^2} \mathbf{w}^\top \mathbf{w}\right) \\ &\propto \exp\left(\sigma^{-2}(\mathbf{y}^\top \mathbf{y} - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w}) + \tau^{-2} \mathbf{w}^\top \mathbf{w}\right) \\ &\propto \exp\left(-2\sigma^{-2} \mathbf{w}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{w}^\top (\sigma^{-2} \mathbf{X}^\top \mathbf{X} + \tau^{-2} \mathbf{I}) \mathbf{w}\right) \end{aligned} \quad (2.5)$$

where we dropped $p(\mathcal{D})$ and all the other terms which were constant with respect to \mathbf{w} . Equation 2.5 has the same form to Equation 2.4 (with respect to \mathbf{w}) and therefore is proportional to a multivariate Gaussian distribution. This relation shows that the posterior should be multivariate Gaussian and thus its parameters can be found by matching the terms of the two equations as $\Sigma^{-1} = (\sigma^{-2} \mathbf{X}^\top \mathbf{X} + \tau^{-2} \mathbf{I})$ and $\Sigma^{-1} \mu = \sigma^{-2} \mathbf{X}^\top \mathbf{y}$. Finally, the posterior of our simple linear regression can be described as

$$\begin{aligned} p(\mathbf{w} | \mathcal{D}) &= N(\mathbf{w} | \mu, \Sigma), \text{ where} \\ \Sigma^{-1} &= (\sigma^{-2} \mathbf{X}^\top \mathbf{X} + \tau^{-2} \mathbf{I}), \\ \mu &= \Sigma \sigma^{-2} \mathbf{X}^\top \mathbf{y}. \end{aligned} \quad (2.6)$$

This simple Bayesian regression is used as the underlying data model in Publication I. It would be interesting to compare a point estimate of the posterior distribution with the frequentist solution $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$. Maximum a posteriori (MAP) is a point estimate which represents the value in the posterior that has the highest density. In our case, this would be equal to the posterior mean, which

³Note that we use $\mathbf{w} \sim N(0, \tau^2 \mathbf{I})$ to denote the random variable \mathbf{w} and $N(\mathbf{w}|0, \tau^2 \mathbf{I})$ to refer to the density function.

after some rearrangement, can be described as $\hat{\mathbf{w}}_{MAP} = (\mathbf{X}^\top \mathbf{X} + \frac{\sigma^2}{\tau^2} \mathbf{I})^{-1} (\mathbf{X}^\top \mathbf{y})$. Comparing the mean of the posterior with the frequentist solution indicates that the two estimates are very similar with the only difference that the MAP estimate has the additive term $\frac{\sigma^2}{\tau^2} \mathbf{I}$ in the covariance matrix. This added term is the direct outcome of our assumptions in the likelihood and prior of the model which were not present in our simple frequentist model. This term can also be interpreted as a regularization parameter that controls the variance of weights. The designer of the model may want to consider different types of regularization or constraints in the model, which in probabilistic modelling, is usually done by incorporating them in the prior distribution. Such problems may not have an analytical posterior available. We will consider two such models in the following subsections.

Sparsity-inducing prior

In many prediction problems, in particular those that require human intervention in a form, the number of available training data is smaller than dimensions of the problem. If not regularized properly, a model trained in this setting may overfit to the training data (achieving very low training error) while not being able to generalize properly to unobserved data (high test data error). One way to tackle this challenge, is to directly regularize the model parameters so that they would not overfit to observations. This regularization can be done by adding penalty terms (for example l_1 norms of the weights) to the error function in the non-Bayesian models (see for example Lasso [5]). In probabilistic modelling, we can achieve the same goal by selecting sparsifying priors.

There are different priors that encourage sparsity but they can be categorized to two general groups of mixture priors, such as spike-and-slab prior [3], and the continuous shrinkage priors (such as horseshoe prior [4]). In this thesis we consider the spike-and-slab prior due to its high interoperability (we discuss more about this in XXXX).

Consider the linear regression model Then explain spike and slab prior and what it means then parameters and posterior then talk a bit about posterior inference (read Juho's paper also and cite Tomi).

This model is used as the underlying data model in Publication II, III, and IV.

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 θ and ϕ_y the model parameters. Let f encode input (*feedback*) from the user, presumably a domain expert, and ϕ_f be model parameters related to the user input. We identify the following key components:

1. An observation model $p(y | x, \theta, \phi_y)$ for y .
2. A feedback model $p(f | \theta, \phi_f)$ 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 θ , 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 θ 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.

TO EXPLAIN: Why we use linear models? because of the risk for overfitting if the model is not regularized enough... and also mentioning what model was used in each publication?

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.
- [3] Edward I George and Robert E McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.
- [4] Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Statist.*, 11(2):5018–5051, 2017.
- [5] Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

Publication I

Pedram Daee, Joel Pyykkö, Dorota Glowacka, 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.

© 2016 ACM

Reprinted with permission.

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.

© 2017

Reprinted with permission.

Publication III

Iiris Sundin*, Tomi Peltola*, Luana Micallef, Homayun Afrabandpey, Marta Soare, Muntasir Mamun Majumder, Pedram Daei, 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.

© 2018

Reprinted with permission.

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.

© 2018 ACM

Reprinted with permission.

Publication V

Giulio Jacucci, Oswald Barral, Pedram Daei, 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.

© 2019

Reprinted with permission.