Abstract

Ongoing title:
Probabilistic user modelling methods for improving human-in-the-loop machine learning
Interactive user modelling for human-in-the-loop machine learning

Abstract

En puhu suomea.

Preface

To fill.

Espoo, September 20, 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 probabilistic 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 [7].

User modelling in human–computer interaction aims at improving the usability and usefulness of collaborative human–computer systems and providing personalised user experiences [9]. Machine learning based interactive systems extend user modelling to encompass statistical models interpreting user's actions.

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. The computational solutions for Bayesian inference are reviewed in 2.4. Finally, the key components for modelling observational data and user feedback as a joint probabilistic model is introduced in Section 2.5.

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 [11]. 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 [11] and use $p(.)$ to denote a probability distribution and $p(. \mid .)$ 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 θ. The observational model for an observation y describes the conditional density of y given the parameter θ and is denoted as $y \sim p(y \mid \theta)$. It is usually assumed that the observations are conditionally independent given θ, 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 θ with fixed observation y. One of the core questions in statistical inference is to estimate the parameter of interest θ based on observations $\mathscr{D}$. Bayesian inference answers this question by computing the conditional distribution of θ 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 θ and $p(\mathscr{D})$ 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., $p(\mathscr{D}) = \int_\theta p(\mathscr{D},\theta)d\theta = \int_\theta p(\mathscr{D} \mid \theta)p(\theta)d\theta$ [1]. $p(\theta \mid \mathscr{D})$ is called the posterior and it expresses the uncertainty surrounding the true value of 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

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

$p(\hat{y} \mid \mathscr{D})$ 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 unobserved, 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

[1] The integral turns to summation for discrete θ.

[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.

the regression tasks, meaning that we model response variables by real values, 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

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

where $\boldsymbol{w} \in \mathbb{R}^d$ is the regression coefficients or the model's weights and ϵ_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 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 expressed in vector format 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 (also known as least squares solution). 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 ϵ_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 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 $\boldsymbol{w}$. 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

$$\begin{aligned} \boldsymbol{y} &\sim \mathrm{N}(\boldsymbol{X}\boldsymbol{w}, \sigma^2 \mathbf{I}), \\ \boldsymbol{w} &\sim \mathrm{N}(0, \tau^2 \mathbf{I}), \end{aligned} \tag{2.3}$$

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 $\boldsymbol{w}$. The Bayesian rule allows us to derive the posterior for $\boldsymbol{w}$ as[3]

$$p(\boldsymbol{w} \mid \mathscr{D}) = \frac{p(\mathscr{D} \mid \boldsymbol{w})p(\boldsymbol{w})}{p(\mathscr{D})} = \frac{\mathrm{N}(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{w}, \sigma^2\mathbf{I})\mathrm{N}(\boldsymbol{w} \mid 0, \tau^2\mathbf{I}))}{p(\mathscr{D})}. \tag{2.4}$$

Generally, the posterior in many problems cannot be analytically derived (we will discuss this issue more in Section 2.4). 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

$$\begin{aligned} \mathrm{N}(\boldsymbol{w} \mid \mu, \Sigma) &\propto \exp\left((\boldsymbol{w}-\mu)^\top \Sigma^{-1}(\boldsymbol{w}-\mu)\right) \\ &\propto \exp\left(\boldsymbol{w}^\top\Sigma^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\top\Sigma^{-1}\mu\right), \end{aligned} \tag{2.5}$$

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

To derive the posterior, we follow Equation 2.4

$$\begin{aligned} p(\boldsymbol{w} \mid \mathscr{D}) &\propto \exp\left(\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})\right)\exp(\frac{1}{2\tau^2}\boldsymbol{w}^\top\boldsymbol{w}) \\ &\propto \exp\left(\sigma^{-2}(\boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}) + \tau^{-2}\boldsymbol{w}^\top\boldsymbol{w}\right) \\ &\propto \exp\left(-2\sigma^{-2}\boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{w}^\top(\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X} + \tau^{-2}\mathbf{I})\boldsymbol{w}\right) \end{aligned} \tag{2.6}$$

where we dropped $p(\mathscr{D})$ and all the other terms which where constant with respect to $\boldsymbol{w}$. Equation 2.6 has the same form to Equation 2.5 (with respect to $\boldsymbol{w}$) and therefore is proportional to a multivariate Gaussian distribution. This relation shows that the posterior should be multivariate Gaussian, as both equations should integrate to one, and thus its parameters can be found by matching the terms of the two equations as $\Sigma^{-1} = (\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X} + \tau^{-2}\mathbf{I})$ and $\Sigma^{-1}\mu = \sigma^{-2}\boldsymbol{X}^\top\boldsymbol{y}$. Finally, the posterior of our simple linear regression can be described as

$$\begin{aligned} p(\boldsymbol{w} \mid \mathscr{D}) &= \mathrm{N}(\boldsymbol{w} \mid \mu, \Sigma), \text{where} \\ \Sigma^{-1} &= (\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X} + \tau^{-2}\mathbf{I}), \\ \mu &= \Sigma\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{y}. \end{aligned} \tag{2.7}$$

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 $\hat{\boldsymbol{w}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$. Maximum a posteriori (MAP) is a point estimate which represents the value in the posterior that has

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

the highest density. In our case, this would be equal to the posterior mean, which after some rearrangement, can be described as $\hat{\boldsymbol{w}}_{MAP} = (\boldsymbol{X}^\top \boldsymbol{X} + \frac{\sigma^2}{\tau^2}\mathbf{I})^{-1}(\boldsymbol{X}^\top \boldsymbol{y})$. 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 constrains 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 priors

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. 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 with the general idea of pushing weights that are not useful toward zero (see for example Lasso [20]). In probabilistic modelling, we can achieve the same goal by selecting sparsity-inducing 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 [12], and the continuous shrinkage priors, such as horseshoe [16] and Laplace prior [18]. The core idea of these priors is that their densities for coefficients peak at zero but at the same time have good amount of probability mass in non-zero values. In this thesis we consider the spike-and-slab prior as it introduces a binary latent variable that explicitly indicates whether a coefficient should be zero or non-zero (relevant and not-relevant groups). This explicit explanation could be very useful as it is compatible with human opinion about inclusion/exclusion of variables in a model (we will discuss this link in detail in Section 2.3).

The linear regression model with sparsity-inducing spike-and-slab prior on coefficients $\boldsymbol{w}$ can be described as

$$\begin{aligned} \boldsymbol{y} &\sim \mathrm{N}(\boldsymbol{X}\boldsymbol{w}, \sigma^2\mathbf{I}), \\ \sigma^{-2} &\sim \mathrm{Gamma}(\alpha_\sigma, \beta_\sigma), \\ w^j &\sim \gamma^j \mathrm{N}(0, \tau^2) + (1-\gamma^j)\delta_0, \qquad j = 1, \ldots, d, \end{aligned} \tag{2.8}$$

$$\gamma^j \sim \text{Bernoulli}(\rho), \qquad j = 1, \ldots, d,$$

where δ_0 is a Dirac delta point mass at zero, and γ^j is the binary variable that indicates whether the corresponding coefficient w^j should be excluded from the model (if $\gamma^j = 0$, then $w^j = 0$) or should it be addressed as a normal variable (if $\gamma^j = 1$, then $w^j \sim \text{N}(0, \tau^2)$). As we are interested in the behavior of γ^js, we considered a Bernouli prior distribution on all of them with the prior inclusion probability ρ as a fixed hyperparameter that controls the expected number of non-zero covariates. Unlike the simple model introduced before, here we consider σ^2 as an unknown parameter and assume a prior distribution for it with α_σ and β_σ as fixed hyperparameters. for the three unobserved parameters $\boldsymbol{w}$, $\boldsymbol{\gamma}$, and σ^2 the posterior can be derived following the Bayes rule

$$p(\boldsymbol{w}, \boldsymbol{\gamma}, \sigma^2 \mid \mathscr{D}) = \frac{p(\mathscr{D} \mid \boldsymbol{w}, \sigma^2) p(\boldsymbol{w} \mid \boldsymbol{\gamma}) p(\sigma^2) p(\boldsymbol{\gamma})}{p(\mathscr{D})}. \tag{2.9}$$

The posterior does not have an analytical solution. We will discuss different methods to compute the posterior of these types of problems in Section 2.4. The introduced spike-and-slab Sparsity-inducing model is used as the underlying data model in Publication II, III, and IV.

Detecting outliers

In regression, there may be observations that have response values substantially different from all other data. These highly noisy observations, called outliers, can considerably affect the results if not accounted properly. A Bayesian way to handle this is to consider different variance for residual error of each observation [21]. This can be implemented by changing the observational model from $y_i \sim \text{N}(\boldsymbol{x}_i^\top \boldsymbol{w}, \sigma^2)$, where there was a global parameter σ^2 for the variance of all observations, to $y_i \sim \text{N}(\boldsymbol{x}_i^\top, \frac{\sigma^2}{\nu_i})$, where the parameter ν_i controls the noise variance per observation. As ν_i is an unobserved parameter, we consider a prior distribution for it to complete the Bayesian loop

$$\begin{aligned}
y_i &\sim \text{N}(\boldsymbol{x}_i^\top \boldsymbol{w}, \frac{\sigma^2}{\nu_i}), && i = 1, \ldots, n, \\
\nu_i &\sim \text{Gamma}(\alpha_\nu, \beta_\nu), && i = 1, \ldots, n, \\
\sigma^{-2} &\sim \text{Gamma}(\alpha_\sigma, \beta_\sigma), \\
\boldsymbol{w} &\sim \text{N}(0, \tau^2 \mathbf{I}),
\end{aligned} \tag{2.10}$$

where α_ν, β_ν, α_σ, β_σ, and τ^2 are fixed hyperparameters and n is the number of observations. For the unobserved parameters $\boldsymbol{w}$, σ^2, and $\boldsymbol{\nu}$ (as a vector of all ν_is), the posterior does not have an analytical solution and can be written as

$$p(\boldsymbol{w}, \boldsymbol{\nu}, \sigma^2 \mid \mathscr{D}) = \frac{p(\mathscr{D} \mid \boldsymbol{w}, \sigma^2, \boldsymbol{\nu}) p(\boldsymbol{w}) p(\sigma^2) p(\boldsymbol{\nu})}{p(\mathscr{D})}. \tag{2.11}$$

Noisy observations can particularly happen if the data provider is a noisy source. For example in Publication V, we considered the setting where some of the response variables were generated from noisy physiological signals of a human user. This model was employed to handle the potential outliers in the data.

2.3 Modelling user interaction for prediction

A core idea in the thesis is to model the user interaction with the probabilistic model as observational feedback. These feedbacks can be tied to the model parameters via conditional distributions (feedback models). The posterior inference can then learn the unobserved parameters given data observations (as introduced in the previous section) and the feedback observations (which will be discussed in this section). The feedback models are used to incorporate the user knowledge into the regression models introduced in the previous section. The following subsections explain the feedback models used in Publication I-V, and their relations to the probabilistic models.

2.3.1 User feedback as observation about model parameters

Publication II and III target the case where the user has knowledge about the regression parameters and can provide feedback if queried. As mentioned, we consider high dimensional scenarios where few data points are available. An example task could be a drug response prediction given genomic features of patients [3]. The genomic feature size may exceed thousands of variables, but only a small number of patient data might be available, which make the prediction task challenging. Fortunately, experts in the field may have experience or knowledge from literature about which features, and in what ways, are relevant for this prediction task. If accounted properly, the expert knowledge can help the prediction performance. Another potential scenario is the sentiment prediction problem [13] when only few texts (e.g., textual reviews of items) with known sentiments (e.g., score of the review) are available. A classical representation of textual data, known as bag-of-words, considers each textual data as a vector of distinct words where the feature value indicates the number of appearance of each word in the text[4]. Humans have intuitions about how individual words may be related to the sentiment (e.g., appearance of the word "awful" in a review may indicate low review score). Our goal is to design probabilistic models that can receive these types of expert knowledge, alongside the training data, to boost the prediction performance.

We consider the linear regression model with sparsity-inducing spike-and-slab prior on coefficients introduced in Equation 2.8 as the underlying data model.

[4]It is also common to represent textual data in dense forms such as [6]. We use bag-of-words due to its simplicity and interpretability of the features.

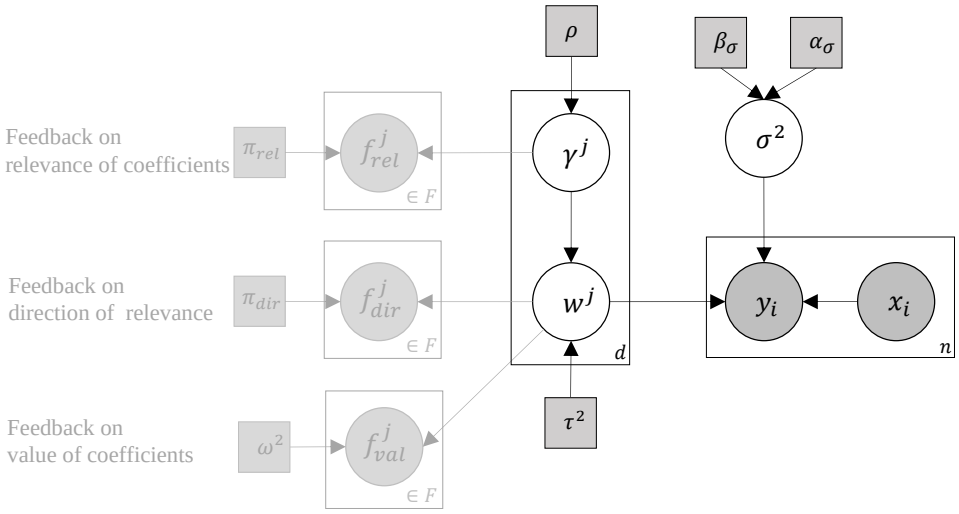

Figure 2.1. Plate notation of the prediction model (right; see Equation 2.8) and user feedback observations (left; see Equations 2.12, 2.13, and 2.14). User feedback influences the data model through observations about model parameters. Circles represent variables (observed variables are shaded), and shaded squares are the fixed hyperparameters.

The type of feedback naturally depends on the task and availability of user knowledge. We consider three simple and natural types of user interaction with the model.

- With some noise, the user can provide feedback about the value of coefficients

$$f^j_{val} \sim \mathrm{N}(w^j, \tau^2), \tag{2.12}$$

  where τ^2 models the variance of error in user feedback. Knowledge about value of coefficients is very powerful, however, only available in some specific applications (e.g., exploiting similar trained models or reported values in related literature).

- With some probability, the user can provide feedback about whether a coefficient should be included or excluded from the model

$$f^j_{rel} \sim \gamma^j \,\mathrm{Bernoulli}(\pi_{rel}) + (1-\gamma^j), \mathrm{Bernoulli}(1-\pi_{rel}). \tag{2.13}$$

  where π_{rel} indicates the probability that user is correct about the feedback. This relevant/not-relevant feedback is the simplest form of knowledge that a user may have about a regression model.

- With some probability, the user can provide feedback about direction of relevance of a coefficient, i.e., whether a feature is positively or negatively correlated with the response variable

$$f^j_{dir} \sim I(w^j \geq 0)\mathrm{Bernoulli}(\pi_{dir}) + I(w^j < 0)\mathrm{Bernoulli}(1-\pi_{dir}), \tag{2.14}$$

  where π_{dir} indicates the probability that user is correct about the feedback.

The connection of the three mentioned feedback with the data model (Equation 2.8) is shown as a plate diagram in Figure 2.1. The full posterior of the unknown parameters given data observations and the user feedback can be described as

$$p(\boldsymbol{w},\boldsymbol{\gamma},\sigma^2 \mid \mathscr{D},\mathscr{F}) = \frac{p(\mathscr{D} \mid \boldsymbol{w},\sigma^2)p(F_{val} \mid \boldsymbol{w})p(F_{rel} \mid \boldsymbol{\gamma})p(F_{dir} \mid \boldsymbol{w})p(\boldsymbol{w} \mid \boldsymbol{\gamma})p(\sigma^2)p(\boldsymbol{\gamma})}{p(\mathscr{D},\mathscr{F})}, \tag{2.15}$$

where F_{val}, F_{rel}, and F_{dir} are the sets of collected feedback corresponding to the three considered feedback types and $\mathscr{F} = (F_{val},F_{rel},F_{dir})$. The posterior computation is discussed in Section 2.4.

Related works
The proposed feedback models provide an intuitive and effortless ways for the user to directly influence the prediction model without caring about the complications of the data model. Classical prior elicitation (see for example [15, 10]) aims at eliciting a distribution to represent the expert's knowledge by asking about summary information such as quantiles of the parameters. This is done through iterations between the expert in the related field and statisticians who design the model. Our work goes beyond pure elicitation as it directly connects the expert to the model and also exploits the training data to facilitate the interaction (will be discussed in Section 3).

Studies have shown that the type of domain knowledge in prediction tasks can be summarize in a small set [4]. A large body of works have studied the exclusion/inclusion of features in different contexts, for classification [17, 8, 19], and regression [14]. These methods are different to ours from the modelling point as we consider sparse models and also the set of possible interactions... Pairwise similarity of features has also been investigated in [1, 2].... Human-in-the-loop feature selection [5].

2.3.2 User feedback as outcome of a cognitive process

Publication IV.

Related works
HERE I NEED TO CITE OUR LAB WORKS REGARDING THIS?

2.3.3 User feedback as data observation

In many tasks only user feedback is available as observations. An immediate instance is the intent modelling in personalised recommender system. Targets of Publication I and V... The model can just be a linear regression. Challenge of limited feedback... In the papers we proposed to augment the type of feedback either by considering new dimension of feedback (e.g., feedback on both keywords and documents) or new type of feedback source (e.g., physiological signal)...

Related works
HERE I NEED TO CITE OUR LAB WORKS REGARDING THIS?

2.4 Posterior inference

2.5 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 $p(y \mid x, \theta, \phi_y)$ 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 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] Homayun Afrabandpey, Tomi Peltola, and Samuel Kaski. Interactive prior elicitation of feature similarities for small sample size prediction. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, UMAP '17, pages 265–269, New York, NY, USA, 2017. ACM.

[2] Homayun Afrabandpey, Tomi Peltola, and Samuel Kaski. Human-in-the-loop active covariance learning for improving prediction in small data sets. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1959–1966. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[3] Muhammad Ammad-ud din, Suleiman A. Khan, Disha Malani, Astrid Murumägi, Olli Kallioniemi, Tero Aittokallio, and Samuel Kaski. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics*, 32(17):i455–i463, 08 2016.

[4] In Kwon Choi, Taylor Childers, Nirmal Kumar Raveendranath, Swati Mishra, Kyle Harris, and Khairi Reda. Concept-driven visual analytics: An exploratory study of model- and hypothesis-based reasoning with visualizations. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 68:1–68:14, New York, NY, USA, 2019. ACM.

[5] Alvaro H. C. Correia and Freddy Lécué. Human-in-the-loop feature selection. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[6] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087, 2015.

[7] 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.

[8] Gregory Druck, Burr Settles, and Andrew McCallum. Active learning by labeling features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 81–90, 2009.

[9] Gerhard Fischer. User modeling in human–computer interaction. *User Modeling and User-Adapted Interaction*, 11(1-2):65–86, 2001.

[10] Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701, 2005.

[11] 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.

[12] Edward I George and Robert E McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[13] Bing Liu. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.

[14] Luana Micallef, Iiris Sundin, Pekka Marttinen, Muhammad Ammad-ud din, Tomi Peltola, Marta Soare, Giulio Jacucci, and Samuel Kaski. Interactive elicitation of knowledge on feature relevance improves predictions in small data sets. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces*, IUI '17, pages 547–552, New York, NY, USA, 2017. ACM.

[15] Anthony O'Hagan, Caitlin E. Buck, Alireza Daneshkhah, J. Richard Eiser, Paul H. Garthwaite, David J. Jenkinson, Jeremy E. Oakley, and Tim Rakow. *Uncertain Judgements. Eliciting Experts' Probabilisties*. Wiley, Chichester, England, 2006.

[16] Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Statist.*, 11(2):5018–5051, 2017.

[17] Hema Raghavan, Omid Madani, and Rosie Jones. Active learning with feedback on features and instances. *Journal of Machine Learning Research*, 7(Aug):1655–1686, 2006.

[18] Matthias W Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008.

[19] Burr Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, 2011.

[20] 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.

[21] J. Ting, A. D'Souza, and S. Schaal. Automatic outlier detection: A bayesian approach. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 2489–2494, April 2007.

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.