Learning Optimal Individual Behavior from Individual/Action Outcome Differences

*

ABSTRACT

Given a group of mutually observing individuals (actions and outcomes), how can each learn his or her optimal strategy? Beyond best-outcome actions, decisions depend on knowledge of the latent differences among individuals in respect to the outcome (How do I differ from others? How do those differences influence action outcomes?) A comprehensive solution requires individuals therefore to model both actions and other group members. We outline this problem, related causality issues and a solution formulated as a metric learning problem.

In the solution, we place actions in a metric space A where distances among action positions are made to reflect expected changes in outcome. We then place individuals in a space B with the same property (i.e., close by individuals are expected to receive similar outcomes to similar actions). This way, an outcome is determined simultaneously by an A-vector (chosen action) and a B-vector (individuals effect heterogeneity). We show how to learn these spaces from observed outcome differences, using background-action-outcome datasets. These datasets contain sets of individual background variables (corresponding to B), action variables (A) and outcome measures ($f \in \mathbb{R}$). We consider the solution's prediction accuracy in examples where individuals choose where to go (e.g., to eat), entrepreneurs choose what new business to start, and policy-makers decide which businesses to incentivize.

CCS CONCEPTS

•**Computer systems organization** → **Embedded systems;** *Redundancy;* Robotics; •**Networks** → Network reliability;

KEYWORDS

Metric-learning, User-modeling, Causality, Behavior Prediction, Economical Analysis, Policy making

1 INTRODUCTION

Choosing a strategy to improve one's performance in a given problem (whether is choosing a date, a school, a new export, or what to watch tonight) is often a hard problem. The difficulty is often related to learning, as choosing a strategy requires estimating an unknown outcome function to be maximized, often with costly sampling. When many individuals face the same problem, their reciprocal attempts can increase individual members' sample sizes (and thus accuracy). This prompts, however, a decision problem for the individual: how (and to what extent) he or she should take observations from others. If not the individuals themselves, policy-makers and institutions face this same problem and must learn how information and solutions affect differently members of their communities (and the practical impact of those differences). Technology is allowing individuals and policy-makers to observe an increasing number of 'others'. While this increase could scale-up, possibly dramatically, the quality of individual solutions, it is not clear how performance, networks and user diversity are related.

The problem appears because predicting outcomes from action-outcome observations is not enough, since outcomes are likely to be different conditional on latent individual characteristics. This means that individuals must not only model actions, but also other individuals. A wide range of techniques (from regression analysis to reinforcement learning and matrix factorization) have been used to learn either the effectiveness of actions (Kearns and Singh, 2002)(Nielsen and Jensen, 2004)(Kober et al., 2013) or the similarity of an individual to others (Goyal et al., 2010)(kumar Bokde et al., 2015)(Beel et al., 2013)(Ricci et al., 2011). We propose an approach that represents these two elements of the problem explicitly and uniformly, learning metrics for actions and individuals simultaneously (from individual-action-outcome observations). Since this requires understanding which action or personal characteristics in fact cause differences in observed outcomes, we explore a combination of causal and metric learning. The metric is learned from outcome differences among every pair of observations and reflect conditional independence among observations' features, both central to characterize a 'causal' effect. We use examples such as difference of average income (*outcome*) among all 70K zipcodes in the US (*individual*) and their businesses (*action*) or difference of review rankings (*outcome*) among 500K consumers (*individual*) and restaurants (*action*). Beyond individual decisions, we are interested in devising a representation to ask and visualize what-if questions about outcomes for all individuals (e.g., all American zip-codes).

The effort leads to two practical contributions. The first is **algorithmic** and introduces an experimental view to metric learning, leading to a metrics that can be used for prediction (next section) and highlights causal relations in the data.

As second contribution, we show how to use the technique to model, under a common framework, the impact of individual and action differences on received outcomes. This suggests the importance of modeling the **two problem elements** (relevant differences among agents and their actions in respect to the outcome) to predict and understand behavior in human systems. The use cases considered make use of both demographic and economic data and introduce new problems for knowledge discovery. The impact of the heterogeneity of subjects (or users) on learning is an important issue often studied in the Social Sciences, and we demonstrate a computational framework to visualize and understand the problem.

*

WOODSTOCK'97, El Paso, Texas USA

DOI: 10.475/123_4

1.1 Problem Formulation

More specifically, consider an unknown outcome function f. An individual i wants to maximize f by choosing an action a (described by a feature set A) but outcomes also depend on individual-specific variables b_i (described by a feature set B of 'background' variables),

$$\arg\max_{a|b_i} f(a|b_i). \tag{1}$$

Notice that individuals can maximize f by either choosing actions directly (a first-order effect) or changing contingent background variables which could change the effect of actions (a second-order, or modulating, effect). We call the first features 'actions' and the second 'background' abstractly. For example, for entrepreneurs or policy makers actions can be what new types of businesses to start or incentivize (where A is a set of businesses types) or for consumers which establishments to visit (A is a set of establishments). In these cases, however, outcomes can also depend on demographic characteristics (B is set of indicators collected in the American population census).

Start with a set of individuals $0 < i, j \leq n$ and a set of actions $0 < k, v \leq m$. And let $O = [a, b, f(a|b)]^c$ be the set of c observations from this group of individuals. Learning in this case can be defined as learning the function $y = f(a|b)$, which individuals can then employ in their maximization. Instead of assuming a prior form for $f(a|b)$, we assume the function is unknown or complex and define it, instead, as a linear combination of group observations,

$$\begin{aligned} &\arg\max_{a_k} f(a_k|b_i) = \\ &\arg\max_{a_k} \Big\{ \sum_{j=0}^{n} (1 - \frac{d^B_{ij}}{v_B}) \sum_{v=0}^{m} (1 - \frac{d^A_{kv}}{v_A}) f(a_v|b_j) \Big\}, \end{aligned} \tag{2}$$

where d^B_{ij} is the distance in the unknown space $\mathcal{B}$ among individuals i and j, v_B is the maximal distance in $\mathcal{B}$ and $(1 - \frac{d^B_{ij}}{v_B})$ is consequently the relevance of an individual's j observation in i's estimation (in respect to f). Similarly, d^A_{ij} defines the dissimilarity of two actions (in respect to f) and v_A is the maximal distance in the implied space $\mathcal{A}$.

These distances correspond therefore to two unknown metric spaces[1], one for individuals, $\mathcal{B} = (O, d^B_{ij})$, and another for actions, $\mathcal{A} = (O, d^A_{ij})$. These two metrics allow the individual to take advantage of *all* group observations (across all individual and all actions they undertook). In this article, we will characterize these metrics and study how to learn them from data.

1.2 Metric Properties

We define two desired properties for the metric. The first property we will investigate is that distances among actions (or individuals subsequently) should reflect differences among outcomes. According to this, two actions with similar effects in f should be close in this metric. The two actions should not, however, have the same position if they have different effects conditional on each other. The second property we will consider is thus that distances among actions should also reflect conditional independence among them when predicting the outcome.

[1] A metric space is an ordered pair (M, d) where M is a set and d is a metric on M, i.e., a function $M \times M \to \mathbb{R}$.

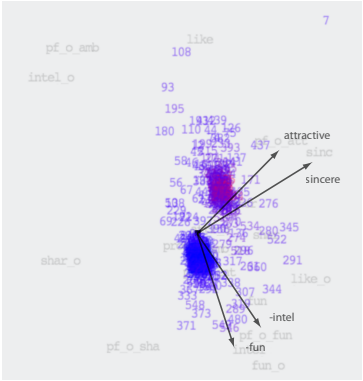

Figure 1: Visualization of Speed dating participants (numbers are participants' unique ids) in a space $\mathcal{A}^{32}_f$ of actions. Color indicates number of accepted propositions a participant received (red is the maximum in the dataset and blue minimum).

The central concept used in graphical models is conditional independence (Pearl, 2000)(Wills, 2002). We will define a distance requirement based on conditional independence that is, additionally, learned from outcome differences among every pair of observation. It provides a continuous analogue to the graphical distance in graphical models (where spatial distance, instead of graphical distance, implies causation and independence). These two properties (the effect of changes and their inter-relationship) will allow us to give a causal interpretation to the representation. The resulting representation reveals different causal 'knobs' in the studied problems. These are low dimension, independent, latent variation ranges with a progressive effect on the outcome f (and consequently, on individual optimization). They (in particular, conditional independence) are also natural requirements for the use of Eq. 2 for outcome prediction. Before discussing these properties in detail, we use a simple example to illustrate and motivate them.

1.3 A Preliminary Example

We suggested that before observations from others can be informative, the individual must learn how he or she differs from them. More specifically, the individual must learn among all possible ways he or she can differ from others which do make a difference on the outcome of interest when choosing actions. This is directly related to the previous properties; to clarify that relationship we use an example.

Consider an everyday problem. In dating, you choose whom to propose next (*action*). The action's outcome (success) correspond to whether the proposition is accepted or not. Here, again, outcome depends on action choices and personal characteristics (such as race, income, etc.). There is therefore not one best strategy, but a personal best strategy. A dater has, as consequence, two types of interventions available. He can choose dates strategically (eg., choose from the prettiest to smartest girl) or change personal characteristics (eg., go to the gym or get a promotion). But his or her

comprehensive optimal behavior is an interaction between personal characteristics and immediate chosen actions. The crux of the problem is that neither the relative effect of these intervention types, nor of each intervention individually, is known a priori.

To make this more concrete, consider a recent dataset on speed dating (Fisman and I., 2006). In speed dating, a group of women and men are paired, each man briefly meeting every woman (and vice-versa). After each meet, they rank each other across 32 factors (e.g., how attractive the other was) and move on to the next meet. At the end of the event, they each mark ('proposition') which others they would go out with. The set of rankings (dataset A) describe the action for the individual (who he proposes).

Define thus an outcome in this case as the number of 'propositions' each individual gets in the end of the event,

$$f_i = 2\frac{(\#\text{ proposals } i \text{ received})}{n}, \tag{3}$$

where n is the number of participants (and there are equal numbers of males and females). In this setting, each individual is placing other individuals they meet in a 32-dimensional space, $\mathbb{R}^{32}$. If we think of each dimension as an attribute that an individual has (or not) and that these attributes fully determine ranking decisions, then for a single individual others are natural 'experiments' on the effect of removing or adding attributes to his or her current attributes. This assumption implies that position differences (i.e., changes of attributes) should hold a relation to outcome differences in the space that represents these choices. This is however not the case for the space $\mathbb{R}^{32}$ (as defined by the researchers' ranking questionnaire). Many dimensions have no effect on the number of propositions a participant get (eg., the station number where met partner). At the same time, for dimensions that have an effect, unit rankings differences have distinct effects on f (eg., unit changes in attractiveness or sincerity are bound to have distinct effects on the outcome).

Fig.1 shows a space $\mathcal{A}_f^{32}$, that, instead, holds this property (a low-dimensional visualization). The space is learned with the data (Fisman and I., 2006) and techniques in Sect.3.3. The numbers in the figure are participants unique ids and are colored based on f (number of proposals) , ranging from red (maximum number of proposals an individual received) to blue (minimum). The norm of vectors in this space correspond to the expected effect differences across individual dimensions. The figure also shows the 4 dimensions with highest expected effect. Changes in these dimensions are expected to change proposal acceptance counts. And the norm of these vectors is proportional to the expected effect of these attributes. In particular, changes in the direction of a particular combination of the four attributes (the red area in Fig.1) should bring the highest increase. We want, additionally, this space to reflect not only correlation (of features when predicting f), but to highlight the causes for changes in f. We then want distances between observations to also reflect their conditional independence when predicting f.

The space $\mathcal{A}_f^{32}$ suggests only actions average effects on the group. Consider however the case where a diverse set of individuals meet (i.e., that have many differences that affect f). In this case, predictions are biased across individuals. To use Eq.(2) to predict maximum gain behavior, a metric over individuals is also necessary.

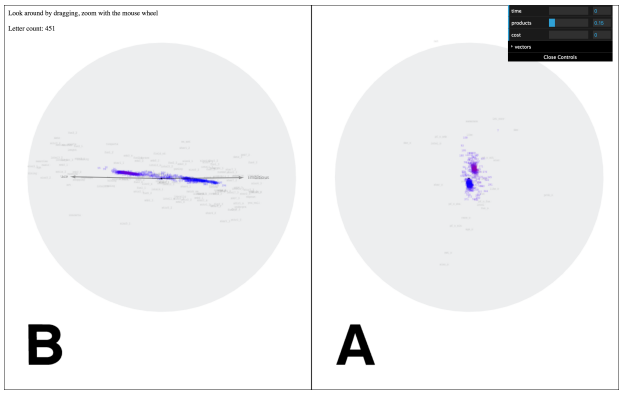

Figure 2: Visualization of Speed dating participants in spaces $\mathcal{A}_f^{32}$ of actions and $\mathcal{B}_f^{151}$ of individuals.

Fig.2 shows instead two maps $(\mathcal{A}_f^{32}, \mathcal{B}_f^{151})$ learned simultaneously: the space $\mathcal{A}_f^{32}$, as before, was learned from individual ranking of others, while the space $\mathcal{B}_f^{151}$ was learned from 151 background variables that describe the individual himself or herself in ways that can potentially affect their rankings (eg., race, income, self-ranking). These variables were also collected by (Fisman and I., 2006). Both spaces reflect the studied relationship between positions ($b \in \mathcal{B}_f^{151}, a \in \mathcal{A}_f^{32}$) and outcome f. The space $\mathcal{B}_f^{151}$ reflects the notion of statistical heterogeneity in social experiments. The issue has been widely studied in the social sciences but rarely appears in computer science research explicitly. We therefore study spaces $\mathcal{A}_f^{32}$ and $\mathcal{B}_f^{151}$ together as a representation of decision problems for individuals and consider to what extent they can predict optimal behavior. We will consider economic datasets that have background variables (demographic or census data) and economic (eg., average neighborhood income) or social media (eg., establishment review rankings) performance measures in Sect.4.

2 BACKGROUND AND RELATED WORK

Distance metric learning is a fundamental problem in data mining and knowledge discovery (Bellet et al., 2013)(Weinberger and Saul, 2009). Informally, the task is to learn a metric that assigns small distances to pairs of observations that are semantically similar (and large to dissimilar observations). Each target problem has its own semantic notion of similarity, which is often badly captured by standard metrics (e.g., Euclidean distance).

In current research the problem is most often formulated for the Generalized Mahalanobis distance (Weinberger and Saul, 2009), which measures the distance between an observation x_i and x_j by

$$d(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j), \tag{4}$$

where M is some arbitrary Symmetric Positive Semi-Definite (SPSD) matrix.

We can see that the Euclidean, Cosine, χ^2 and Mahalanobis distances can directly be computed given the data, i.e., no learning procedure is needed. Distance metric learning , in contrast, is necessary when we want to learn the best precision matrix M from the data such that some optimality criterion is met. Unsupervised distance metric learning methods do not require any supervision, i.e., they learn an optimal distance metric purely from the data matrix of observations, such that some geometric or discriminative optimality is achieved. PCA (Jolliffe, 2011) is a popular example, and can be interpreted as learning a matrix M aiming at extracting the projection directions from the data on which the maximum variance is achieved.

More generally and closer to the view in this article, it is possible to see the metric learning problem with a Vector space point-of-view. For a pair of data points x_i and x_j, the problem is now to learn a mapping function f, such that $f(x_i)$ and $f(x_j)$ will be in the Euclidean space and $d(x_i, x_j) = |f(x_i) - f(x_j)|$, where $|.|$ is the $l2$- norm.

With this view, word-embedding methods can be seen as metric learning solutions. These methods have proven extremely practical in Natural Language modeling. Mikilov (Mikolov et al., 2013) introduced word2vec, a novel word-embedding procedure. His model learns a vector representation for each word using a (shallow) neural network language model. Specifically, he proposes a neural network architecture (the skip-gram model) that consists of an input layer, a projection layer, and an output layer to predict nearby words. Each word vector is trained to maximize the log probability of neighboring words in a corpus, i.e., given a sequence of words $w_1, ..., w_T$,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j\in nb(t)} logp(w_j|w_t), \tag{5}$$

where $nb(t)$ is the set of neighboring words of word w_t and $p(w_j|w_t)$ is the hierarchical softmax of the associated words, see (Mikolov et al., 2013) for more details. Several other embeddings are also plausible (Collobert and Weston, 2008) (Mnih et al., 2009) (Pennington et al., 2014) Shi and Liu (2014) and have been proposed after word2vec.

We refer to these as regression approaches to metric learning, as they attempt to explicitly regress a metric from pairs of observations and correlation/frequency data. Due to its surprisingly simple architecture and the use of the hierarchical softmax, the skip-gram model can be trained on a single machine on billions of words per hour using a conventional desktop computer. The ability to train on very large data sets allows the model to learn complex word relationships such as $vec(Einstein) - vec(scientist) + vec(Picasso) \approx vec(painter)$ (Mikolov et al., 2013). Learning the word embedding is entirely unsupervised and it can be computed on the text corpus of interest or be pre-computed in advance.

We propose a specific weighted least squares regression model that trains, instead, on pairs of outcome differences (and not word frequencies) and uses a different, causally-inspired, objective function. In Sect.4, we consider whether the model produces a vector space with meaningful substructure and/or predictive power.

3 LEARNING NORMS AND DISTANCES

We first formulate a relationship between vector norms (in action $\mathcal{A}$ or individual $\mathcal{B}$ spaces), vector distances and their effect in the outcome f. We then consider how to learn both these spaces simultaneously.

3.1 Modeling Effect and Effect Heterogeneity

Given that we are interested in modeling human behavior, we start with how modeling the effect of interventions in human systems is most commonly formulated in Economics and the Social Sciences, where the issue has been studied extensively. Consider a researcher wishes to estimate the effect of an action i on an outcome of interest. He or she begins with with observed data in the form of values $(\mathbb{I}_i, f_i)^n$, where f is an observed outcome variable and $\mathbb{I}_i$ is a treatment status indicator variable (which indicates whether a participant performed an action - e.g., received a treatment or intervention). The goal is to determine the causal effect of $\mathbb{I}_i$ on f. The notation often adopted is

$$\begin{aligned} f_i &= \mathbb{I}_i f^1 + (1 - \mathbb{I}_i) f^0 \\ &= f^0 + (f^1 - f^0)\mathbb{I}_i \end{aligned} \tag{6}$$

where f^1 and f^0 are average outcomes of those who did and did not perform the action. This says that the action effect is determined by the difference of average outcomes between those two groups.

Assume next that participants can additionally be described by high-dimensional covariate vectors, x_i^B, for each participant i. A common next step is to first regress (eg., linearly) f on vectors x_i^B and then on the indicator $\mathbb{I}_i$,

$$f_i - f^0 = \delta\mathbb{I}_i + (x_i^B\beta) + \varepsilon, \tag{7}$$

where δ is the 'effect' of action k, β is a vector of predictor weights, f_0 is a bias/intercept term and ε is an error term. This requires finding the maximum-likelihood vector β across all participants' covariates and observations $(x_i^B, f_i)^n$. It is possible to give this formulation a causal interpretation (Morgan, 2007).

Our goal for the next sections is to review this problem (and Eq.7 in particular) with a vector-based representation. We first assume differences in outcome can be explained by differences in individual positions in an unknown action space $b_i \in \mathcal{B}$. In particular, we define a relationship between a participant i's vector norm and projection on other participants' vectors. For the case of only two participants, i and j, the solution will have the form

$$f_i - f_j = \delta\mathbb{I}^i + (|b_i| - b_i * b_j) + \varepsilon, \tag{8}$$

where $*$ is the dot product of two vectors. This is further justified in the next sections. We then extend the formulation to the case of many actions (not only i). We assume as result that individuals and actions can be described by two high-dimensional vectors, a_i and b_i,

$$f_i - f_j = (|a_i| - a_i * a_j) + (|b_i| - b_i * b_j) + \varepsilon. \tag{9}$$

This says, in contrast to Eq.7, that outcome differences are determined by distances among actions and individuals in two spaces $\mathcal{A}$ and $\mathcal{B}$. It requires finding Maximum-Likelihood spaces $\mathcal{A}$ and $\mathcal{B}$ across all pairs (i, j) of covariates and observations $[(a_i, b_i, f_i), (a_j, b_j, f_j)]$. Because the representation is learned from pairs of observations, metric learning is a natural framework to study this problem. We focus on the pairwise differences case. The extension to higher order relationships (triplets, quadruplets, etc.) is possible by exploring independence in a way similar to Bayesian Networks (Wills, 2002).

3.2 Set-Based Representation

We first consider only actions (and the space $\mathcal{A}$), and assume no heterogeneity (all individuals obtain the same effect, given they take the same actions). In the next sections, we extend the solution, first, to encompass uncertainty in the performance of an action (leading to a vector representation) and, second, to both vector spaces ($\mathcal{A}$ and $\mathcal{B}$).

Let $A_k = \{a_1, ..., a_n\}$ be a set of actions with additive effects, performed by an individual k. The set is therefore a many-actions generalization of the indicator function $\mathbb{I}_i$ in the last section. The individual has attributes A_k and the strategic question revolves around which *other* attributes to acquire. Consider then observed action sets from two other individuals, A_i and A_j, and the problem of reaching the decision, based on these observations, whether the optimal action set for k is $A_i \cup A_k$ or $A_j \cup A_k$. That is, should the individual k choose to acquire elements from the set A_i or A_j? Considering these two distinct hypothesis about the optimality of action-sets, the Neyman-Person Lemma (Lehmann and Romano, 2005) suggests that a decision could be based on the likelihood ratio. These are not hypotheses, however, about the effect of set A_i or A_j, but about the the sets $A_i - A_k$ (the actions that i took but not k) and $A_j - A_k$. We define the distance between action A_k and A_v (for all k) as

$$d_{ij} = \frac{\sum_{x \in A_i - A_k} f(x)}{\sum_{x \in A_j - A_k} f(x)} = \frac{f_i - f_k}{f_j - f_k} \tag{10}$$

This defines equally effective (but distinct) sets of actions at similar distances to k. Sets at same distance provide this way alternative paths for the individual to take. We can also demonstrate that this reflects conditional independence between the two action-sets. Let $D_f(A_i, A_j|A_k)$ be the conditional dependence of A_i on A_j when predicting f (given A_k). We estimate dependence with the ratio between the expected value of f conditional only on A_i and on both A_i and A_j. This should take a value of 1 if the two sets are indepedent. We want the distance between i and j to reflect the ratio, in turn, between dependences (for all k). With the no-heterogeneity assumptions, we formulate the distance between i and j in respect to every third single individual k as

$$\begin{aligned} d_{ij} &= \frac{D_f(A_i, A_j|A_k)}{D_f(A_j, A_i|A_k)} \\ &= \frac{\mathbf{E}[f|A_i - A_k]}{\mathbf{E}[f|A_i - A_k + A_j - A_k]} \left(\frac{\mathbf{E}[f|A_j - A_k]}{\mathbf{E}[f|A_i - A_k + A_j - A_k]} \right)^{-1} \\ &= \frac{\mathbf{E}[f|A_i - A_k]}{\mathbf{E}[f|A_j - A_k]} = \frac{f_i - f_k}{f_j - f_k} \end{aligned} \tag{11}$$

With a continuous space, it is natural to assume an exponentially decaying function for each observed f (centered at the observed position). This leads to the likelihood ratio of these observations,

$$d_{ij} = \frac{\exp(\sum_{x \in A_i - A_j} f(x))}{\exp(\sum_{x \in A_j - A_i} f(x))} \tag{12}$$

3.3 Vector-Based Representation

We make the presence of an attribute stochastic next, letting each take values in $[0, 1]$ (resp. from not present to present). Actions sets now correspond to vectors, and the distance in Eq.12 becomes,

$$d_{ij} = \frac{\exp(|a_i| - a_i * a_k)}{\exp(|a_j| - a_j * a_k)}. \tag{13}$$

Notice that set difference is not captured by vector difference (eg., pairwise element subtraction can be negative), but by the projection of the vector a_k in a_i or a_j.

We can now turn to how to learn this metric from every pair of observations. Consider again the relationship between two vectors a_i, a_j, and a third vector a_k. According to Eq.13, the position of a_i should reflect the following ratio for all k,

$$\sum_i \sum_k \frac{\exp(|a_i| - (a_i * a_k))}{\exp(|a_j| - (a_j * a_k))} = \frac{(|f_i - f_k|)}{(|f_j - f_k|)} \tag{14}$$

If these constraints are applied to every pair of actions, i and j, then by reflexibility we consider the fraction terms to be independent objectives. We then minimize the *log* of the difference between rhs and lhs in Eq.3.3 for each action,

$$\arg\min_{a_i} \sum_i \sum_k |a_i| - (a_i * a_k) - log(|f_i - f_k|) \tag{15}$$

This single-position objective is extended and implemented for multiple actions in a way typical to vector embedding approaches (Mikolov et al., 2013), but where words correspond to actions and the objective above is used instead. As result, this takes the initial action values and minimizes the above objective with a gradient descent (for all actions simultaneously). The initial value for each coordinate (i.e., likelihood of having performed a given action) are given empirically by the datasets studied.

This introduces a different objective function and input to regression-based metric learning approaches that use, instead, conditional probability ratios as distance metrics (Mikolov et al., 2013) (Pennington et al., 2014) (Shi and Liu, 2014). We will next extend further the approach to learn both $\mathcal{A}$ and $\mathcal{B}$ spaces in parallel.

3.4 Spaces $\mathcal{A}$ and $\mathcal{B}$

The issue of causality takes central place in the design or analysis of experiments, since experiments typically try to validate causal relations. Informally, a 'good' experiment is one where one causal factor is changed while possible others are kept constant or are accounted for in the data. Factors can be accounted by design (e.g., by randomizing on possible factors) or by analysis (e.g., by conditioning-out possible factors' effects). More formally, in the experimental framework discussed in Sect.3.1, a set of variables is said to fully stratify x_i^B and $\mathbb{I}_i$ if $x_i^B \perp \mathbb{I}_i$ (Morgan, 2007). This concept appears not only in Statistics, but in Computer science - corresponding to Pearl's back-door criterion in Bayesian Networks (Pearl, 2000) - and Economy - to controlled variable sets (Angrist, 2009).

We can look at the problem with the previous framework and explore a relationship between experimental design and metric learning. Consider the outcome difference between any two actions as an observation for learning a metric. An observation where two similar individuals perform two actions are experimental opportunities to learn an action metric. Reversely, observations where two similar actions are performed are opportunities to learn a metrics over individuals.

We will therefore learn two metrics, for spaces $\mathcal{A}$ and $\mathcal{B}$, in parallel. In line with the previous discussion, we will employ a weighted regression for each metric, where weights when learning metric $\mathcal{A}$ are given by (the reciprocal of distances) in space $\mathcal{B}$, and vice-versa,

$$\begin{aligned}
&\arg\min_{a_i} \sum_i \sum_k (1 - \frac{d_{ik}^B}{v_B})^\beta \Big[|a_i| - (a_i * a_k) - log(|f_i - f_k|) \Big], \\
&\arg\min_{b_i} \sum_i \sum_k (1 - \frac{d_{ik}^A}{v_A})^\beta \Big[|b_i| - (b_i * b_k) - log(|f_i - f_k|) \Big],
\end{aligned} \tag{16}$$

where v_A and v_B are maximum distances between two actions or individuals in the two metric spaces. We additionally assume the weight function is exponential with the distance, described by a power-law. A parameter β thus captures how to weight these contributions. A value of $\beta = 1.0$ weights all observations equally. A small value of β takes into consideration only very similar individuals when learning differences among actions. The parameter thus captures a property of the problem, in particular, how contingent predicting f is (how much differences affect the outcome) given the observed samples.

For the results in this article and the action space, the action dataset A provides the initial positions and the background dataset B provides the (Euclidian) distance in the other space (and, reciprocally, for the background space). We thus perform both regressions weighted by the unlearned distances from the other datasets. A more sophisticated, iterative procedure is however also possible.

3.5 Complexity and Implementation

From Eq.16, the complexity of the model depends on the number of dissimilar outcome observations. The upper bound of the model complexity is $O(|O|^2)$. Due to the nature of the datasets used, complexity is often below the bound. For example, review rankings are used as outcome below and they are an aggregation of 5-level discrete rankings, leading to many same-f observations. In all results, we additionally apply a threshold of 0.01 (to what outcome differences are considered zero) that further sparsifies the regression.

As mentioned, feature values for all vectors, from datasets A and B, are normalized to the maximum observed value in each feature, mapping them to the interval $[0, 1]$ (an estimate of the presence of that attribute). And we do the same for outcomes, also mapping them to the interval $[0, 1]$.

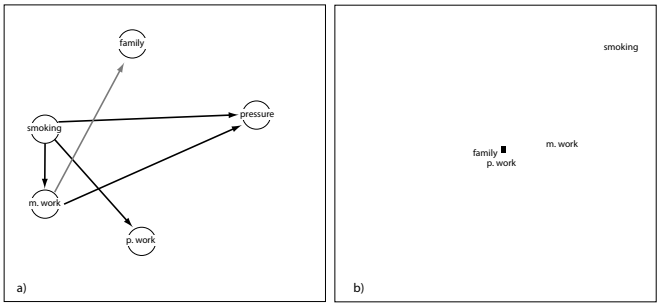

Figure 3: (a) Causal network learned from two methods (see text), (b) Vector-based representation with variable 'pressure' as outcome.

4 RESULTS

4.1 The Relationship to Causal Network Structure Learning

Before considering prediction tasks, we highlight the relationship between the proposed representation and causal networks (Pearl, 2000). In causal networks, a directed edge denotes a causal relation between the connecting variables. Consider the popular 'coronary' dataset that describes the relationship between smoking, work patterns and medical conditions (Wills, 2002). It contains the following variables,

- *smoking*: a two-level factor with levels no and yes.
- *m. work*: strenuous mental work, a two-level factor with levels no and yes.
- *p. work*: strenuous physical work, a two-level factor with levels no and yes.
- *pressure*: systolic blood pressure, a two-level factor with levels <140 and >140.
- *family*: family anamnesis of coronary heart disease, a two-level factor with levels neg and pos.

Suppose we have patients' blood pressure as the health indicator and outcome of interest. There are many proposed algorithms to derive the causal network structure from data like this. As example, Fig.3a shows the output of the max-min hill climbing algorithm (Tsamardinos et al., 2006), a greedy algorithm. The gray arrow shows the only network difference between the output of hill-climbing and the the Markov blanket algorithm of (Yaramakala and Margaritis, 2005).

Fig.3b shows a space $\mathcal{A}$ trained with the methods above (and no space $\mathcal{B}$). The two representations capture similar information. They both indicate that smoking has a possible causal effect on blood pressure. The network indicates the effect with the presence of an edge, while the vector representation with the norm of the vector for *smoking*. Both representations also indicate the effect of mental work and no effects for other variables. The angular differences in the vector representation thus indicate two possibly independent ways to intervene on blood pressure: mental work and smoking. The relationship between *family* and *p.work* is indicated as not strongly independent. The difference of output among the two network learning algorithms can be understood with the vector representation, which doesn't rely on thresholding to indicate their independence.

4.2 Predicting Economic Success

We use the previous framework to address the following problem. Let variables $\mathcal{B}$ be demographic variables (eg., number of individuals of Hispanic origin, with high-school degrees, etc.), $\mathcal{A}$ be businesses types and f be average income of individual US zip-codes. These datasets contain data about individuals or businesses, but aggregated at the level of zip-codes. The resultant problem abstracts the decision problem for policy-makers, where changes in local economic patterns or demographic patterns can possibly lead to economic growth (in this case estimated with the average income of citizens).

Furthermore, we are not only interested in understanding how to choose interventions in these zipcodes but how a particular zipcode currently differs from others and for what possible (economic or demographic) reasons. Stressing causal relations in this dataset is therefore important.

Such count data of demographic and economic variables for all zip codes in the USA is publicly available from the US Census Bureau (Plottel, 2016). Demographic and economic data correspond respectively to the American Community Survey (*individual*) (ACS) and the American Business Patterns (*action*) (ABP) programs. These counts allow us to calculate attribute membership probabilities by simply normalizing values from the maximum observed value. We use average income in a zip-code as the *outcome* variable. It corresponds directly to an ACS variable, and we manually removed any variables related to income from the ACS dataset. For the counts of Businesses types, we also use two zip-code level alternative datasets: business in Google places and a private datasets of credit card purchases. We initially report results using the Google dataset, then consider differences among these datasets.

The problem we want to solve in this case is to predict which interventions (in demographic or economic features) lead to maximal economical growth. While we cannot perform these interventions, we use the census across-time or across-space data as 'natural experiments' to that end. We train a range of models to predict f then use them to predict out-of-sample (or future) observations given new economical and demographic information.

Before considering this prediction task, reconsider, for this dataset, why learning a metric is necessary. Many of the almost 9K attributes in the Census ACS (and their differences in value) can be expected to have no 'causal' effect on f. More specifically,

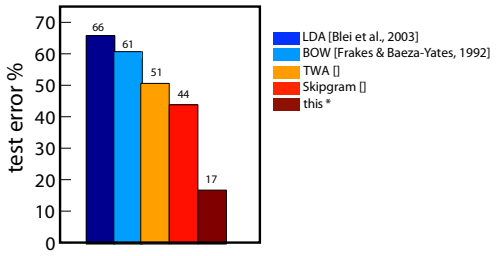

Figure 4: Alternative frequency-based metrics used for prediction with Eq.2.

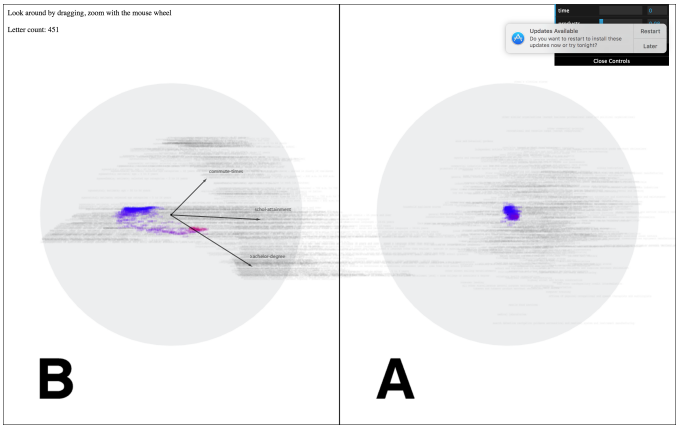

Figure 5: Visualization Of Census ACS and ABP datasets.

different attributes differences are expected to have distinct effects in f and a unit of change in one attribute does not correspond to a unit of change in f.

The low-dimensional visualization in Fig.5 illustrates these different cases more directly. Each number is a zip code, colored according to outcome (average zipcode income, with red as maximum and blue minimum in the dataset). It shows additionally, in gray, several census variables descriptions, corresponding to the spaces dimensions. The norm of these variables are related to their effect on the outcome. Around 87% of dimensions have little effect and small norm (clustering around the origin) and were omitted. Distance in the space is related to similarity in effect. The visualization thus also suggests that many of the remaining variables are redundant and carry similar information when predicting income. Variables in the ACS are grouped in tables of similar demographic indicators. The figure shows the 3 tables with largest sum of norms (vectors in Fig.5's $\mathcal{A}$ space). They correspond to the tables for school attainment, bachelor degree for population and commute time to work.

We start by taking action and background features as a single set of features in a standard prediction test. We use the Census 2011 data as training sets. For tasks using the Census ABP (that has across time data), we generated 50 testing datasets by random sampling 30% of all zip codes in 2011 and evaluating prediction accuracy of outcomes (average income) for 2015. For other tasks,

Table 1: Popular Regression Approaches in joint Economic (A) and Demographic (B) variables, classified in Linear (L), Non-Linear (NL) and Non-Linear with Decision Trees. See text and attached scripts for details.

	Model	F1
L	GLM	73.3 ± 0.9
	MULTINOMIAL	74.2 ± 0.4
NL	MDA	74.6 ± 0.5
	SVM	77.3 ± 3.9
	KNN, Euclidean	76.1 ± 0.6
	NBAYES	69.2 ± 0.7
	NNET	73.5 ± 4.1
	FDA	73.0 ± 1.0
NLT	BAG-CART	77.8 ± 4.0
	BOOST-C50	77.8 ± 4.2
	PART	77.4 ± 4.5
	CART	63.5 ± 7.1
	C45	0.66 ± 2.5
	RF	76.0 ± 2.7
	JRIP	75.0 ± 4.0
	H2O	75.2 ± 1.2

we generate again 50 testing datasets by using 20% of other random zipcodes as testing sets. Table 1 shows results for a range of popular learning solutions. It shows 20 Models (Hastie, 2001) grouped sequentially in 3 groups for the Google places dataset: L) Linear, NL) Non-linear and NLT) Non-Linear Regression with decision trees. The solutions are Multinomial (MULTINOMIAL, L), Logistic Regression (GLM, L), Mixture Discriminant Analysis (MDA, NL), Feed-Forward Neural Net. (NNET, NL), Flexible Discriminant Analysis (FDA, NL), Support Vector Machine (SVM, NL), k-Nearest Neighbors (KNN, NL), Naive Bayes (NBAYES, NL), Classification and Regression Trees (CART, NLT), C4.5 (C45, NLT), C4 PART (PART, NLT), C5 Bagging CART (BAG-CART, NLT), C6 Random Forest (RF, NLT), C9 JRip (JRIP, NLT) and C10 H2O Deep Learning (H2O, NLT).

From the collected Confusion Matrix performance, we compare prediction accuracies. We use popular and open-source R packages. For reproducibility, R-scripts for all solutions were placed in the anonymous repository https://github.com/cikm1724/cikm1724. Most implementations are non-parametric or have embedded parameter optimization modules. For the few exceptions, please see the scripts.

We set out to study whether a metric over outcome differences could be predictive, using Eq.2. To illustrate this point, Fig.4 shows the use of the same equation, with 4 different learned metrics over datasets. Each metric (for each separate dataset A and B) was learned in an unsupervised fashion. We then used Eq.2 to predict outcomes and report test errors (with the same test protocol as before). As discussed in Sect.2, these largely make use of frequency and conditional frequency of observed features. We consider the following baselines, bag-of-words (BOW) model, LDA, Skip-Gram and TWE. The BOW model represents each feature as a bag of

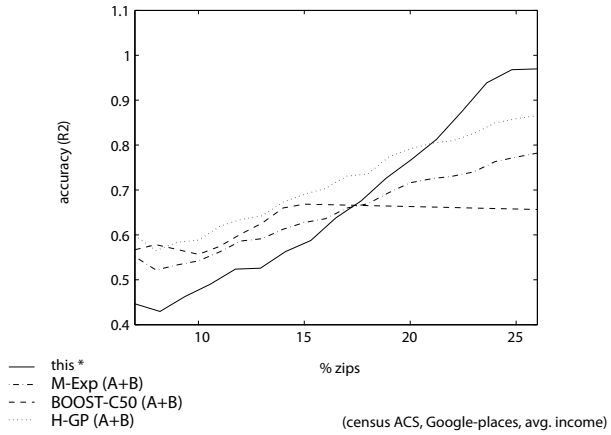

Figure 6: Hierarchical methods and training subsets.

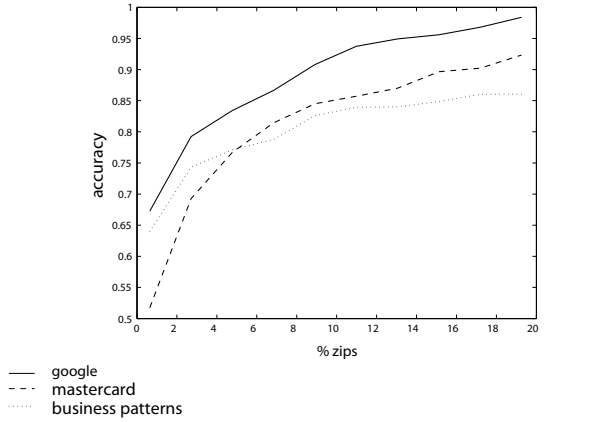

Figure 7: Alternative A-datasets using the current solution and increasing random training subset.

words and the weighting scheme is TFIDF. For the TFIDF method, we select top 50,000 words according to TFIDF scores as features. LDA represents each observation as its inferred topic distribution. In Skip-Gram, we build the embedding vector of an observation by simply averaging over all word embedding vectors in this observation. The dimension of word embeddings in Skip-Gram is also K = 400.

We consider next if making a conceptual distinction between background and action features have an effect in the results. We train accordingly three hierarchical models on the same datasets. This derives, in opposition, a separate model for individuals, a model for actions and a model to combine the output of the last two. These solutions are therefore closer to the proposed solution in this article. And thus we look at these solutions in further detail.

Mixture of experts (ME) is one of the most popular and interesting combining methods, which has great potential to improve

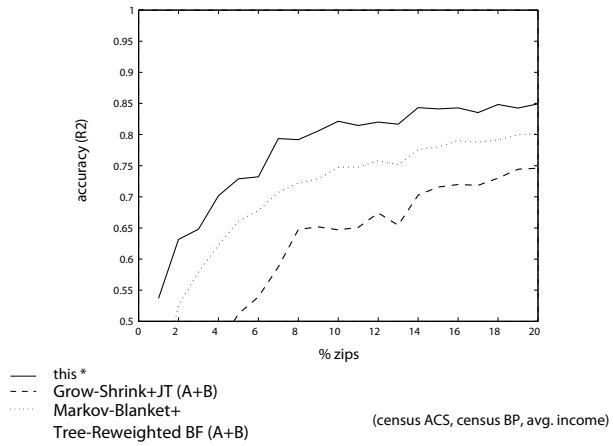

Figure 8: Causal Models over Census ACS and ABP datasets.

performance in machine learning. ME is established based on the divide-and-conquer principle in which the problem space is divided between a few neural network experts, supervised by a gating network. In earlier works on ME, different strategies were developed to divide the problem space between the experts. We use the EM-based implementation of (Tipping and Bishop, 1999).

The next method divides up the input space by making binary splits on the value of a single feature so that partition boundaries are parallel to coordinate axes. In particular, it uses a gradient descent to train multiple Classification and Regression Trees (CART) (Breiman et al., 1984) with boosting. The last model is the Gaussian process latent variable model (GP-LVM), a powerful approach for probabilistic modelling of high dimensional data through dimensional reduction, extended to hierarchies. A hierarchical model (a tree, in this case) allows the model to express conditional independencies in the data as well as the manifold structure. Fig.6 shows prediction accuracy in the Google places A dataset for these approaches with increasing randomized sample sizes. It illustrates that the proposed method outperforms more complex solutions, conditional on data availability. In this case, the method requires a sample size of approximately 13K zipcodes to surpass other solutions' prediction accuracy.

The Census-ABP dataset (an action dataset *A*) provide counts of business types as classified by the North American Industry Classification System (NAICS) (Vogel and (Firm), 2001). NAICS is the standard used by Federal statistical agencies in classifying business establishments. To demonstrate the sensibility of this solution to the classification, Fig.7 shows results with business data from two alternative sources. The first is the set of business counts in individual zipcodes according to the Goggle Places API and the second is a non-public dataset with the number of credit-card purchases per business-type.

Finally, we consider how the current solution compares to Bayesian networks. We used two different algorithms to learn the network structure and two to make predictions. The first two were a Grow-Shrink method and a Markov Blanket algorithm (Sebastiani, 2015). The inference algorithms were a Junction-tree algorithm and a Tree-reweigthed Belief-propagation (Nagarajan, 2013). Fig.8 shows the two best performing combinations using the Census-BP action dataset.

Table 2: Review prediction alternative models accuracy and recall.

Model	Acc.	Prec.	Rec	F1
H-GP	79.7	79.5	79.0	79.0
BOOST-C50	72.2	70.8	70.7	70.0
M-Exp	75.4	75.1	74.3	74.2
this*	82.6	82.5	81.9	81.2

4.3 Predicting User Reviews

We repeat the previous task with a different dataset. Instead of a measure of economic success, we consider the measure of quality implied by user reviews. Like before, we assume outcomes are generated by individual and businesses pairings. Consequently, an individual review is likely to depend simultaneously on the reviewers' personal characteristics (e.g., cultural heritage, income, race) and the ranked businesses.

In this case the outcome variable corresponds to Google places user reviews (a 5-point average ranking of an establishment) and actions are chosen establishments (all places in a zipcode, according to Google maps). Examples are restaurants, hair salons, car shops, doctors etc. Since reviews are often related to information retrieval, recall is also a relevant indicator (in combination with accuracy). We repeat the same experimental procedure as before with the Google places A dataset (30% of random-sampled US zipcodes as training set). We report in Table 2 macro-averaging precision, recall and F1-measure for comparison. The proposed approach score both higher accuracy and recall compared to the three studied hierarchical methods (Gaussian Process, Mixture of Experts and Boosted CARTs).

5 CONCLUSION

We demonstrated a new approach to model human-systems. The solution is comprehensive and makes use of both individual and action large-scale descriptive data. First results suggest the solution can be both interpretable and predictive. In particular, the approach proposes a framework in which effect heterogeneity can be estimated, visualized and used to predict optimal behavior for individuals. We plan to use the tools developed here to reconsider the nature of information in communities and, in particular, the relationship among community diversity and performance.

REFERENCES

Joshua David Angrist. 2009. *Mostly harmless econometrics : an empiricist's companion.* Princeton. Includes bibliographical references (p. 339]-359) and index.; ID: http://id.lib.harvard.edu/aleph/011813983/catalog.

Joeran Beel, Stefan Langer, Andreas Nrnberger, and Marcel Genzmehr. 2013. *The Impact of Demographics (Age and Gender) and Other User-Characteristics on Evaluating Recommender*

Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, 396–400. https://doi.org/10.1007/978-3-642-40501-3_45 ID: Beel2013.

Aurlien Bellet, Amaury Habrard, and Marc Sebban. 2013. A Survey on Metric Learning for Feature Vectors and Structured Data. (2013).

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. (2008), 160-167 pages. https://doi.org/10.1145/1390156.1390177

R. Fisman and I. 2006. Gender differences in mate selection : evidence from s apeed dating experiment. *The quarterly journal of economics* 121, 2 (2006), 673–698. https://doi.org/10.1162/qjec.2006.121.2.673

Amit Goyal, Francesco Bonchi, and Laks Lakshmanan. 2010. Learning influence probabilities in social networks. (2010), 241-250 pages. https://doi.org/10.1145/1718487.1718518

Trevor Hastie. 2001. *The elements of statistical learning : data mining, inference, and prediction*. New York. Includes bibliographical references (p. 509]-522) and indexes.; ID: http://id.lib.harvard.edu/aleph/009036099/catalog.

Ian Jolliffe. 2011. Principal Component Analysis. (2011), 1094-1096 pages.

Michael Kearns and Satinder Singh. 2002. Near- Optimal Reinforcement Learning in Polynomial Time. *Machine Learning* 49, 2 (2002), 209–232. https://doi.org/1017984413808

Jens Kober, J. A. Bagnell, and Jan Peters. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research* 32, 11 (2013), 1238–1274. https://doi.org/10.1177/0278364913495721

Dheeraj kumar Bokde, Sheetal Girase, and Debajyoti Mukhopadhyay. 2015. Role of Matrix Factorization Model in Collaborative Filtering Algorithm: A Survey. (2015).

E. Lehmann and Joseph Romano. 2005. *Testing Statistical Hypotheses*. New York, NY. https://doi.org/10.1007/0-387-27605-X

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. (2013).

Andriy Mnih, Zhang Yuecheng, and Geoffrey Hinton. 2009. Improving a statistical language model through non-linear prediction. *Neurocomputing* 72, 7 (2009), 1414–1418. https://doi.org/10.1016/j.neucom.2008.12.025

Stephen L (Stephen Lawrence) Morgan. 2007. *Counterfactuals and causal inference : methods and principles for social research*. New York. Includes bibliographical references (p. 291-316) and index.; ID: http://id.lib.harvard.edu/aleph/010910135/catalog.

Radhakrishnan Nagarajan. 2013. *Bayesian Networks in R : with Applications in Systems Biology*. New York, NY. ID: http://id.lib.harvard.edu/aleph/013672628/catalog.

Thomas D. Nielsen and Finn V. Jensen. 2004. Learning a decision maker's utility function from (possibly) inconsistent behavior. *Artificial Intelligence* 160, 1 (2004), 53–78. https://doi.org/10.1016/j.artint.2004.08.003

Judea Pearl. 2000. *Causality : models, reasoning, and inference*. Cambridge, U.K. ; New York. Includes bibliographical references (p. 359-373) and indexes.; ID: http://id.lib.harvard.edu/aleph/008372583/catalog.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

Tina Plottel. 2016. United States Census. (May 26, 2016). http://libguides.gwu.edu/USCensus

Francesco Ricci, Lior Rokach, Bracha Shapira, and B. Paul. 2011. *Recommender Systems Handbook*. Boston, MA. https://doi.org/10.1007/978-0-387-85820-3

Paola Sebastiani. 2015. Bayesian Networks: With Examples in R. (2015), 880 pages.

Tianze Shi and Zhiyuan Liu. 2014. Linking GloVe with word2vec. (2014).

Michael E. Tipping and Christopher M. Bishop. 1999. Mixtures of Probabilistic Principal Component Analyzers. *Neural computation* 11, 2 (1999), 443–482. https://doi.org/10.1162/089976699300016728

Ioannis Tsamardinos, Laura Brown, and Constantin Aliferis. 2006. The max- min hill- climbing Bayesian network structure learning algorithm. *Machine Learning* 65, 1 (2006), 31–78. https://doi.org/10.1007/s10994-006-6889-7

Scott M. Vogel and Harris InfoSource (Firm). 2001. *Harris' complete guide to NAICS : your ultimate reference to NAICS, SIC ISIC codes*. Twinsburg, Ohio. ID: http://id.lib.harvard.edu/aleph/009309423/catalog.

K.Q. Weinberger and L.K. Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research* 10 (2009), 207–244.

Graham J. Wills. 2002. Introduction to Graphical Modelling. (2002), 197 pages. https://doi.org/10.1198/tech.2002.s722

S. Yaramakala and D. Margaritis. 2005. Speculative Markov blanket discovery for optimal feature selection. (2005), 4 pp. pages. https://doi.org/10.1109/ICDM.2005.134