ECON484 Machine Learning

5. Dimension Reduction (Part 1, Feature Selection)

Lecturer:
Bora GÜNGÖREN
bora.gungoren@atilim.edu.tr

Concept of a Latent Variable

- Consider a classification problem with N items and 2 classes. Any solution is among a pre-known (albeit large, ie. 2^N) alternatives.
  - Although any classification method will reach a conclusion, its explanatory power will still be limited due to the limited search space created by design choices.
  - Having K-variables in each item will not change this limitation because based on the individual values of these K-variables, one will simply encode the K-variations into the 2 classes.

Concept of a Latent Variable

- So, **why did we select these K variables** specifically?
  - They were in the data set which we have no command on.
  - They were easy to measure and/or measurements were reliable.
  - They were selected on some expert's or the machine learning model builder's own experience (ie. bias).
- A core idea is that there could be some **latent variables**, **unobserved** and therefore unrecorded yet **having explanatory power** on **why** our classes were formed this way.
  - If we had these variables in our data set, we would probably have a much better model, so one could also say **because we missed out on the latent variables is a reason explaining current error levels**.

Concept of a Latent Variable

- So what practical value does latent variables have?
  - **Figuring out the existence of latent variables and measuring the effect of their being missed out** can result in **further exploratory studies** to understand the cause and effect relationships in our problem domain.
    - Example. Diagnosing diseases. If the explanatory power of our diagnosis is low for a set of existing conditions, there should be an additional disease.
  - Using the same approach (inside-out) we can **reduce collinearity problems** or we can **separate variables that were originally blended in our model**, and improve the model.
    - This is much easier to do because there is no need for further data collection.
    - Therefore this is the common approach.

Concept of a Latent Variable

- Example. In psychiatry, there is a manual called DSM-V, which classifies mental disorders and their symptoms. DSM-V is based on several decades of observations made on past patients, and peer-reviewed by experts.
  - Therefore, DSM-V is very similar to a machine learning classifier.
  - It is organized in three forms: textual descriptions, decision trees and (symptom – diagnosis matching) tables.
  - If it behaves like a ML classifier and also looks like one, is it one?
  - Not necessarily, because psychiatrists still receive years of training, to move from a preliminary (provisionary) diagnosis to a differential diagnosis.
  - However DSM-V will serve its purpose in discussion of latent variables.

Concept of a Latent Variable

- Differential Diagnosis in psychiatry (according to DSM-V) works in 6 steps.
- All steps have the **purpose of ruling out a type of latent variable**, which were historically **discovered when trying to understand why psychiatrists made mistakes in their diagnosis**.
  - Rule Out Malingering and Factitious Disorder. A doctor should determine whether patients are faking their symptoms or not.
    - Malingering Disorder is when people feel they have something to gain from a particular diagnosis. For example, they may want to avoid certain responsibilities.
    - Factitious Disorder is when people derive psychological benefits from taking the role of a sick person.
  - Rule Out Drug-Related Causes
    - In many cases drugs interfere with the psyche of a person, so that what is observed is not due to a mental disorder.

Concept of a Latent Variable

- Example. Differential Diagnosis in psychiatry (according to DSM-V)
  - Rule Out General Medical Conditions
    - For example diabetes symptoms are often confused for depression.
  - Determine the Primary Disorder
  - Differentiate It From Other Categories
    - "Other" indicates that a person has a **cluster of symptoms that don't presently exist as a discrete diagnostic category**
    - "Unspecified" indicates that **a person's symptoms don't neatly fit into an existing category**. However, with more information, a diagnosis may be possible.
  - Establish Boundary
    - Does this qualify as a mental disorder or not? (ie. match with good confidence level to a category)

Factor Analysis (FA)

- Factor analysis tries to model N observed k-dimensional vectors, x_i , by describing each data point with a smaller set of f < k unmeasured (latent) variables $z_{i,}$. These variables are called **factors**.
  - This works on defining the **conditional probability** $P(x_i \mid z_i, \Theta)$, where Θ is representing many variables including z_i.
  - We also assume that x_i is a linear combination of z_i, so that **there exists a matrix W (factor loading matrix)** which can be used to **transform z_i to x_i**, ie. $x_i = Wz_i$.
  - We also assume that **z_i explains the covariance in x_i completely**. This is a big assumption, but makes the mathematical operations much simpler (and calculations faster).

Factor Analysis (FA)

- Many techniques are variations of FA, making further assumptions in the math-work.
  - For example assume that some factors in z_i are actually in x_i.
    - This means the transformation matrix W has some rows in the form of many zeroes and a 1, so that one among z_i matches exactly to one of the x_i.
    - Trying to calculate cov (x_i) = $WW^T + \psi$ where ψ is the covariance z_i much easier because the term WW^T (illegible).
  - Another example, Because x_i are actually linear combinations of other x_i, some are also nearly (why not exactly?) linear combinations of their fellow x_i. Can we assume that?
    - If we assume, then our standard statistical toolbox would be useful and enough in just eliminating these variables and reducing the problem space.
    - But the latent variables would still not be identified.

Feature Selection vs Feature Projection (Extraction)

- To minimize the effects of noise, correlation, and high dimensionality, some form of dimension reduction is sometimes a desirable pre-processing step for machine learning.
- Feature selection and extraction are two approaches to dimension reduction.
  - Feature selection: Selecting the **most relevant** attributes
  - Feature extraction: **Combining attributes** into a new reduced set of features
- **Automated tools** exist for this kind of work, but one should be aware of **which** techniques these tools employ and **when**.
  - Example: Oracle's Data Mining software.
    - Uses **feature selection** for optimization within the built-in **Decision Tree** algorithm and within **Naive Bayes** when the user sets **Automatic Data Preparation (ADP)** option as enabled. The pre-processing is done using **Generalized Linear Model (GLM)** method.
    - **Feature extraction** is however a manual step, where the user has to choose between available methods: **Explicit Semantic Analysis (ESA)**, **Non-Negative Matrix Factorization (NMF)** which is default, **Singular Value Decomposition (SVD)**, and **Principle Component Analysis (PCA)**.

Feature Selection

- Feature selection is primarily focused on removing redundant or non-informative predictors from the model.
- Three categories of feature selection
  - Filter methods (ANOVA, Pearson correlation, variance thresholding)
  - Wrapper methods (forward, backward, and stepwise selection)
  - Embedded methods (Lasso, Ridge, Decision Tree).
- In feature selection we usually call x_i as the **predictors** and out model result (ie. classes) as **outcomes**.
  - Note that this is a selection approach, so z_i are assumed to be among x_i.

Feature Selection

Filter methods	Wrapper methods	Embedded methods
Generic set of methods which do not incorporate a **specific machine learning algorithm.**	Evaluates on a **specific machine learning algorithm** to find optimal features.	Embeds (fix) features during **model building process.** Feature selection is done by observing each iteration of model training phase.
Much **faster** compared to Wrapper methods in terms of time complexity	**High computation time** for a dataset with many features	Sits **between Filter methods and Wrapper methods** in terms of time complexity
Less prone to **over-fitting**	High chances of **over-fitting** because it involves training of machine learning models with different combination of features	Generally used to reduce **over-fitting** by **penalizing** the coefficients of a model being too large.
Examples – **Correlation, Chi-Square test, ANOVA, Information gain** etc.	Examples - **Forward Selection, Backward elimination, Stepwise selection** etc.	Examples - **LASSO, Elastic Net, Ridge Regression** etc.

Feature Selection

- Filter methods use statistical calculation to evaluate the relevance of the predictors outside of the predictive models and keep only the predictors that pass some criterion.
  - Considerations when choosing filter methods are the types of data involved, both in predictors and outcome, (ie. numerical or categorical).
- Some notes:
  - We use a measure other than error rate to determine whether a specific feature is useful.
  - A subset of the features is selected through ranking them by a useful descriptive measure (chosen by the particular method).
  - Easy to compute quickly.
  - Will never result in over-fitting.

Feature Selection

- Three filter methods
  - **ANOVA** (Analysis of variance) test looks a the variation within the treatments of a feature and also between the treatments.
    - If the variance within each specific treatment is larger than the variation between the treatments, then the feature hasn’t done a good job of accounting for the variation in the dependent variable.
    - We use the F-test primarily.
  - **Pearson correlation coefficient** is a measure of the similarity of two features that ranges between -1 and 1. So we use the **absolute value** of the coefficient.
    - We need a cut-off point to stop selecting features. Typically we use 0.7 but it is not a fixed rule.
    - Heat maps are used a lot to visualize the selection process.
  - In **variance thresholding**, we focus on the variance of a feature which determines how much predictive power it contains.
    - Given this fact, variance thresholding is done by finding the variance of each feature, and then dropping all of the features below a certain variance threshold.
    - The typical threshold is 0.5

Feature Selection

- In small scale models, drawing a Dendogram diagram and a heatmap will also visualize what typical statistical methods will achieve.
- However when there are many features, a Dendogram loses its easy to use explanatory power. In the image below, the study starts with **4.295** genes.

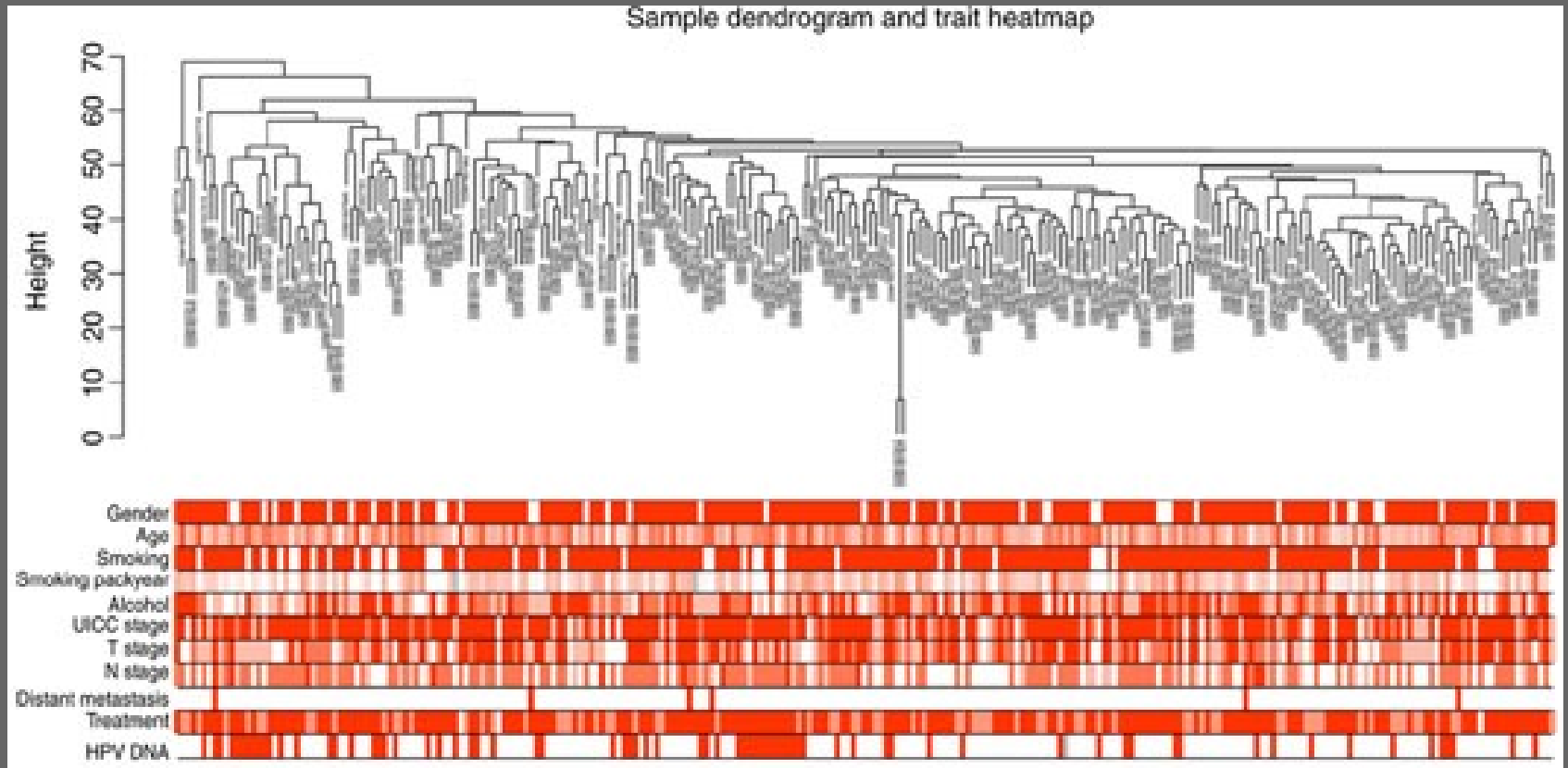

Figure 2. Sample cluster dendrogram and trait indicators. T, primary tumor; N, lymph node; UICC, Union for International Cancer Control; HPV, human papillomavirus.

Feature Selection

- Wrapping methods compute models with a certain subset of features and evaluate the importance of each feature.
  - Then they iterate and try a different subset of features until the optimal subset is reached.
  - This is not easy to do computationally, and takes a lot of time.
  - Also it requires a large data set to do reliably. So you should make a comparison of your sample size N, and the feature size k.
  - If not done properly, can result in over-fitting in your machine learning model.
  - The advantage is, once you select the features in your training data set, this selection is final, and your actual (more complex) model runs on smaller data.
- Wrapping methods can be classified as **model pre-tuning**.

Feature Selection

- Three approaches in wrapper methods
  - **Forward selection** starts with one predictor and adds more iteratively.
    - Start with zero features, then, for each individual feature, run a model and determine the p-value associated with the t-test or F-test performed. Then select the feature with the lowest p-value and add to the working model.
    - Therefore at each iteration, the best of the remaining original predictors are added based on performance criteria.
  - **Backward elimination** starts with all predictors and eliminates one-by-one iteratively.
    - In short, the **feature with the largest insignificant p-value will then be removed** from the model, and the process starts again.
    - A very popular algorithm is **Recursive Feature Elimination (RFE).**
  - Step-wise selection is bi-directional, based on a combination of forward selection and backward elimination.
    - It does reconsider adding predictors back into the model that has been removed (and vice versa).
  - All three approaches may get stuck at a local optima.

Feature Selection

- Embedded approaches conduct feature selection as part of the model tuning.
- Three approaches in embedded methods
  - Ridge does not remove features, but prevents a dominating feature.
  - Lasso may remove features
  - Decision trees work differently (will be covered later).

Feature Selection in R and Python

- In R we use mlbench and caret packages.
  - The manual for caret explains each alternative technique in detail - https://bit.ly/3DMNhsR
  - There is also
- In Python, we use most typical of libraries
  - For filter methods, we use only sklearn. A short tutorial - https://bit.ly/373m4pU
  - For wrapper methods, we use mlextend and sklearn. A short tutorial – https://bit.ly/3JiZXZD

Questions?

CONTACT:
bora.gungoren@atilim.edu.tr