A General Framework for Addressing "Any" Machine Learning Problem

Darren Homrighausen

Visiting Assistant Professor, Dept. of Statistical Science, Southern Methodist University

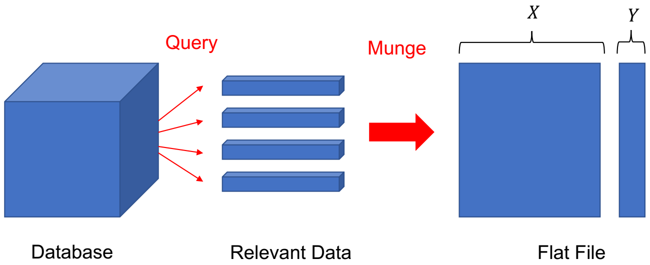
September 21, 2017

The Setup

A machine learning problem can be broken up into two parts:

1. Querying and then cleaning and/or manipulating the data into a format suitable for analysis

   (Sometimes referred to as munging)



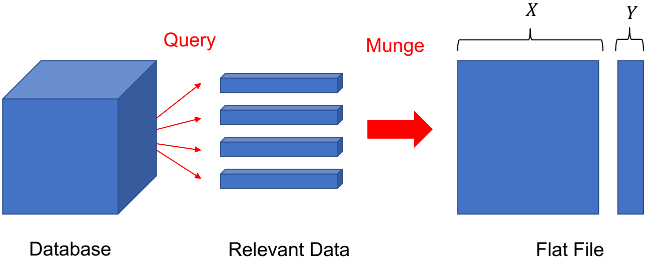
2. Applying machine learning methods to the data

The Setup

A machine learning problem can be broken up into two parts:

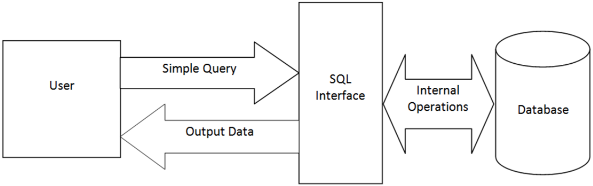
1. Querying and then cleaning and/or manipulating the data into a format suitable for analysis

   (Sometimes referred to as munging)



2. Applying machine learning methods to the data

Querying



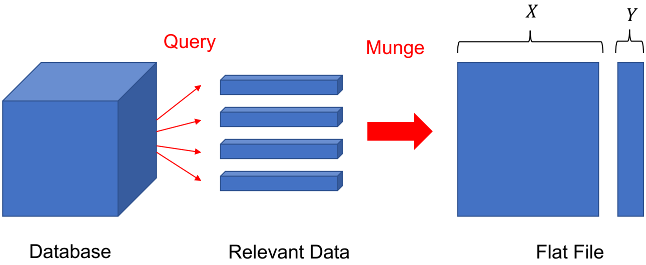

```
INSERT INTO interestingData
SELECT id, trans, city, date
  FROM cust_table
  WHERE date > 1/1/2015
  ORDER BY city, date;
  .
  .
  .
```

The Setup

A machine learning problem can be broken up into two parts:

1. Querying and then cleaning and/or manipulating the data into a format suitable for analysis

   (Sometimes referred to as munging)



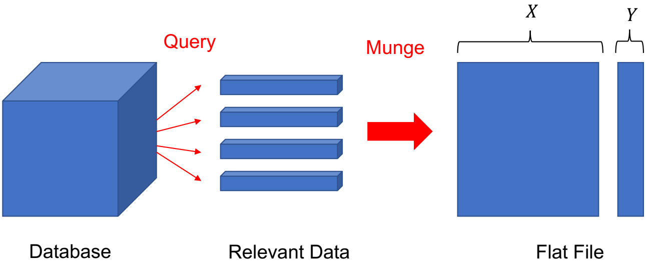
2. Applying machine learning methods to the data

The Setup

A machine learning problem can be broken up into two parts:

1. Querying and then cleaning and/or manipulating the data into a format suitable for analysis

   (Sometimes referred to as munging)



2. Applying machine learning methods to the data

The features

We need to determine the...

... appropriate processing of X

(Known as the features or inputs)

Example:

X: vectorized sky survey

Semi-supervised learning for photometric supernova classification★

Joseph W. Richards,[1,2]† Darren Homrighausen,[3] Peter E. Freeman,[3] Chad M. Schafer[3] and Dovi Poznanski[1,4]

Features

- Qualitative
  - Dummy vars
- Quantitative
  - Scaling?
- Text
  - Text transformer
- Missing
  - Impute

The features: Qualitative

```
         x1  x2
1 -0.6264538  no
2  0.1836433 yes
3 -0.8356286 yes
4  1.5952808  no
```

Gets transformed to...

```
         x1 x2no x2yes
1 -0.6264538    1     0
2  0.1836433    0     1
3 -0.8356286    0     1
4  1.5952808    1     0
```

The features: Quantitative

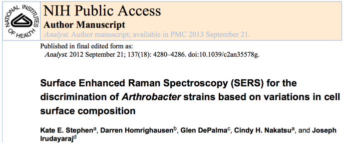
Many methods are not invariant to scale

The usual way of addressing this is...

... Do standardize all features for which scale is meaningful:

$$X \leftarrow \frac{(X - \text{mean}(X))}{\text{sd}(X)}$$

... Don't standardize any
scale-free nor sparse features


NIH Public Access
Author Manuscript
Analyst. Author manuscript; available in PMC 2013 September 21.

Published in final edited form as:
Analyst. 2012 September 21; 137(18): 4280–4286. doi:10.1039/c2an35578g.

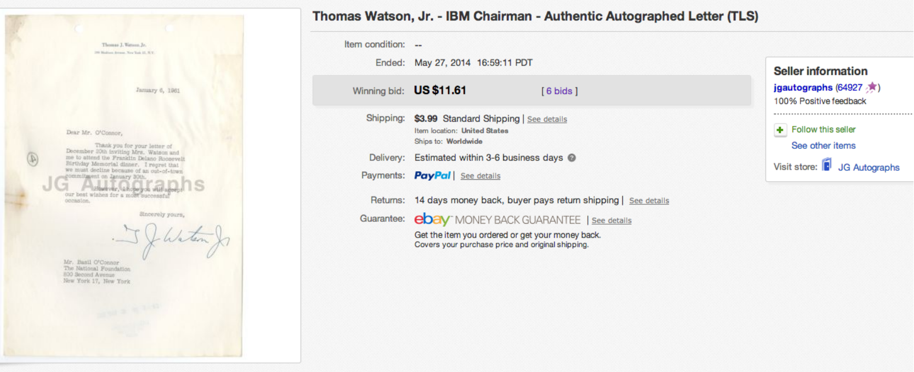
Surface Enhanced Raman Spectroscopy (SERS) for the discrimination of *Arthrobacter* strains based on variations in cell surface composition

Kate E. Stephen[a], Darren Homrighausen[b], Glen DePalma[c], Cindy H. Nakatsu[a], and Joseph Irudayaraj[d]

$$\underbrace{\begin{matrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}}_{X}$$

(Care must be taken if normalizing sparse data)

The features: Text



The features: Text

Buyer:

Always a pleasure! Smooth & pleasant transaction!	f***a (3618)	Jun-10-14 13:52
Thomas Watson, Jr. - IBM Chairman - Authentic Autographed Letter (TLS) (#390846670600)	US $11.61	View Item

Seller:

Great communication. A pleasure to do business with.	Buyer: f***a (3618)	Jun-05-14 18:59
Thomas Watson, Jr. - IBM Chairman - Authentic Autographed Letter (TLS) (#390846670600)	--	View Item

The X matrix can then be written as $X = \begin{bmatrix} X_1^\top \\ X_2^\top \\ \vdots \end{bmatrix}$ where...

$$\begin{array}{lccccccc} & \texttt{always} & \texttt{pleas} & \texttt{smooth} & \texttt{transact} & \texttt{great} & \texttt{commun} & \texttt{busin} \\ X_1^\top = & [\,1 & 2 & 1 & 1 & 0 & 0 & 0\,] \\ X_2^\top = & [\,0 & 1 & 0 & 0 & 1 & 1 & 1\,] \end{array}$$

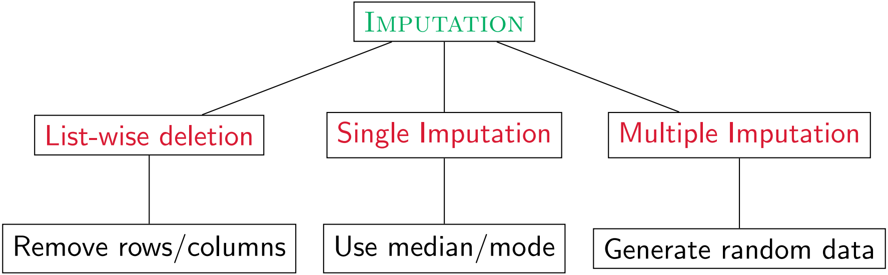
Darren Homrighausen darrenho@stat.colostate.edu
Department of Statistics, Colorado State University, Fort Collins, CO 80523
Gregory Ellison gellison@stat.colostate.edu
Department of Statistics, Colorado State University, Fort Collins, CO 80523

The features: Missing

Corrupted, unrecorded, or unreliable data is commonly referred to as missing data

In statistics, correcting for missing data is known as imputation

There are many, many techniques available:



THE FEATURES: MISSING

- Data size/complexity
  (Does it fit in RAM?)
- Business purpose
  (Is data precious? Development time?)
- Are any observations/features missing a large fraction of values?
- Type of features
  (Any sparsity? Is multivariate normality appropriate?)
- Any atypical missing value indicators?
  (e.g. using -1000 for income to indicate a missing value)

The supervisor

We need to determine the...

... nature of Y

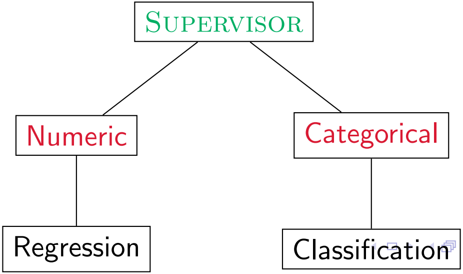
(Known as the supervisor(s) or output(s))

Example:

$$Y = \begin{cases} 1 & \text{If type 1a supernova} \\ -1 & \text{If not} \end{cases}$$

Semi-supervised learning for photometric supernova classification*

Joseph W. Richards,[1,2]† Darren Homrighausen,[3] Peter E. Freeman,[3] Chad M. Schafer[3] and Dovi Poznanski[1,4]



Evaluation Metrics

How to judge success?

Often, this is just mean square error or miss-classification rate

There can be many others:

Example: When classifying supernovae, it is bad to incorrectly label a Type Ia supernova

$\rightarrow$ Evaluation metric:

$$\left(\frac{1}{\text{Total }\#}\right)\frac{(\#\text{ Correctly labeled})^2}{\#\text{ Correctly labeled} + 3(\#\text{ Incorrectly labeled})}$$

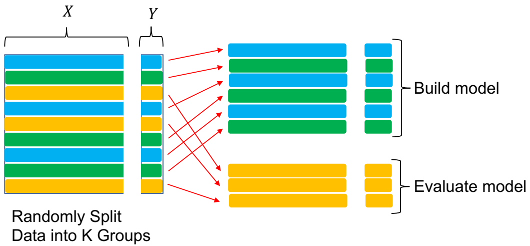
Validation set

We need a realistic measure of the evaluation metric

If at all possible, set aside a (random) validation set

(Say, 10% of the data)

Alternatively (or additionally) a common approach is K-Fold Cross-Validation (CV):



Very important: Make sure to use stratified sampling over

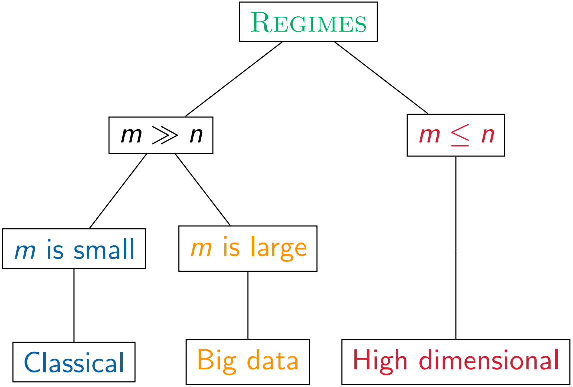
- Any rare, qualitative features
- The supervisor Y

  (If doing classification)

The analysis

Turning these ideas into procedures

There are roughly three regimes of interest, assuming $X \in \mathbb{R}^{m \times n}$



Additionally: Is the data sparse?

(i.e. Does it have a lot of zeros?)

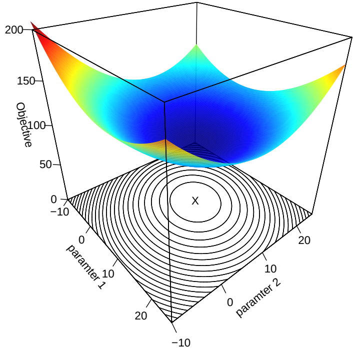
Big data

Big data is usually characterized by 4 "V's"

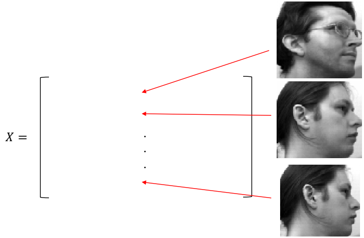
(Volume, Variety, Velocity, Veracity)

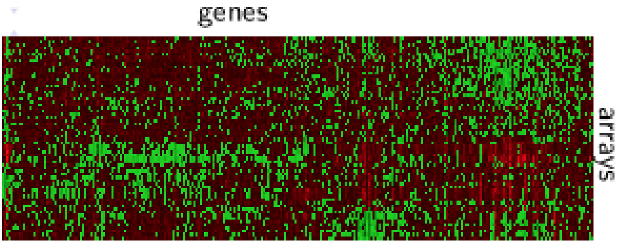
Depending on the data and the desired method, we could:

- Combine randomized projections together with in-memory procedures
- Use stochastic gradient descent (or related methods)
- Leverage an iterative implementation for exact computation (e.g. the QR decomposition for least squares)
- Break the computations down into small bits and distribute these to different cores/processors/nodes (e.g. using the MapReduce paradigm)



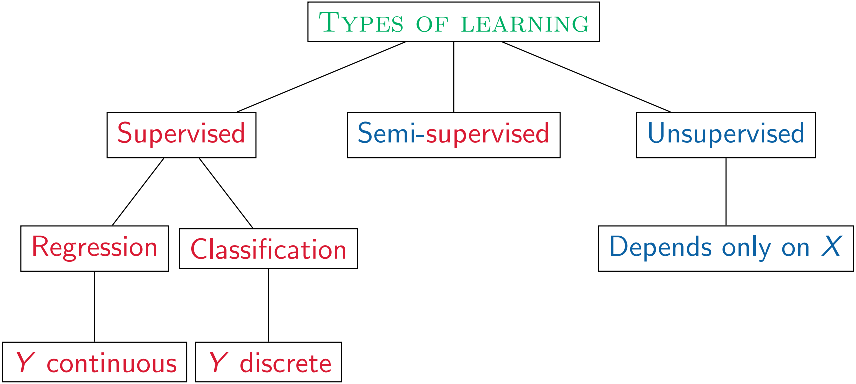
High dimensional regime: Examples





Methods

Methods: Types of learning



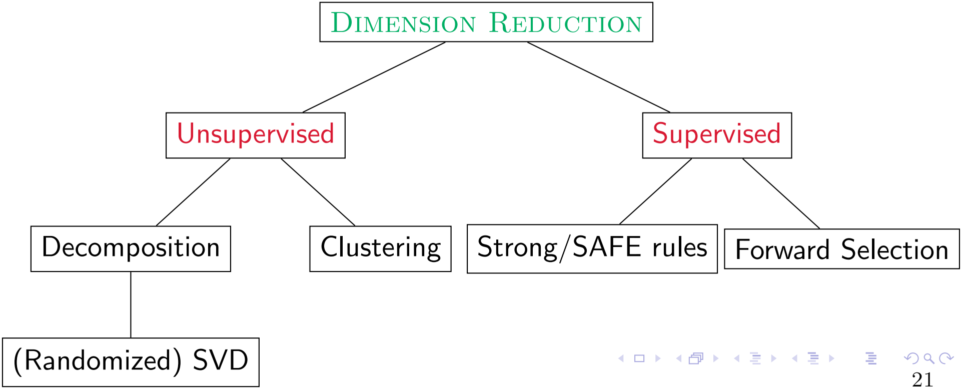
Methods: Dimension reduction

Dimension reduction can help by ...

... reducing the computational load at later steps

... improving prediction performance

Some examples..



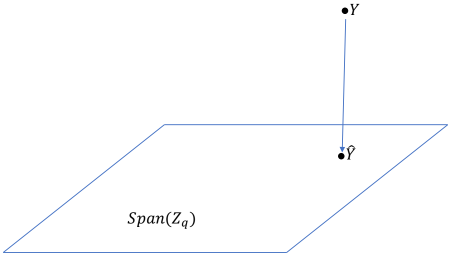
Methods: Supervised

- **Classification:**
  - (Sparse) Logistic Regression
    (I include Linear Discriminant Analysis (LDA) here)
  - Naive Bayes
  - Support Vector Machines (SVM)
  - k-Nearest Neighbors (KNN)
- **Regression:**
  - (Sparse) Linear Regression
    (I include Elastic Net here)
  - Support Vector Regression
- **Both:**
  - Random Forest
  - Gradient Boosting Machines (GBM)
  - Neural Networks

Example: Semi-supervised learning

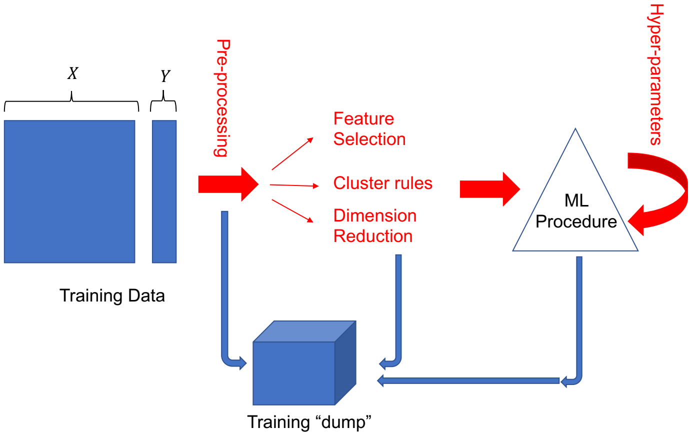
1. Form $X = UDV^\top$

   (That is, the singular value decomposition of the (scaled) matrix X)
2. Project Y onto the column space spanned by the first q columns of UD (call this object Z_q)



(This is commonly referred to as "principal components regression")

Analysis flowchart



- It is very important to save any preprocessing/transformations. These must be applied to the test features
- Choose hyper-parameters via CV or other risk estimator

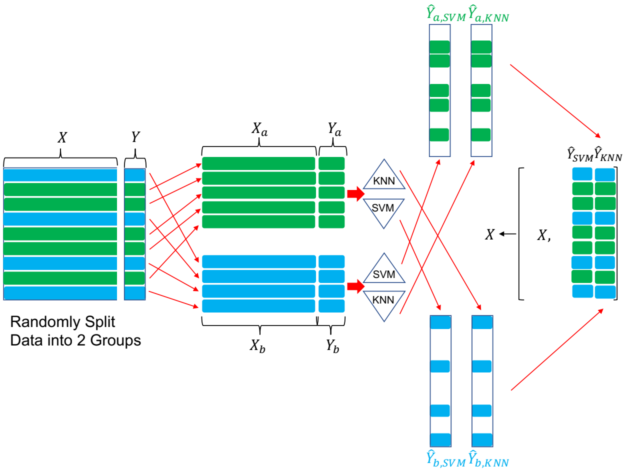
(e.g. q from principal components regression)

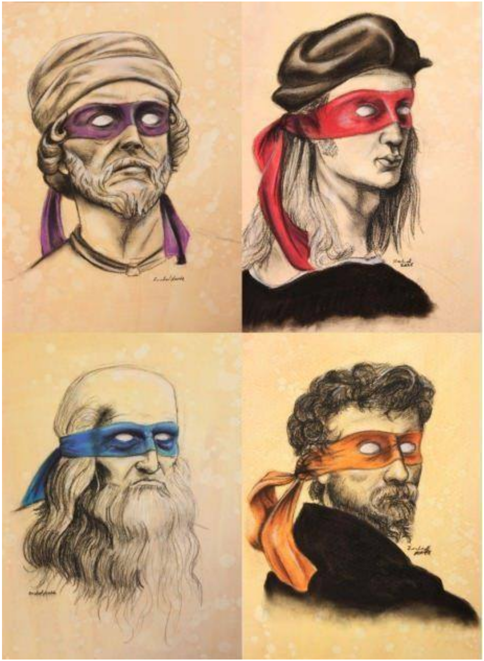
Ensemble methods

Combining supervised methods can result in improved performance

These ensembles or stacks can be formed in several ways

A feature based approach is as follows:

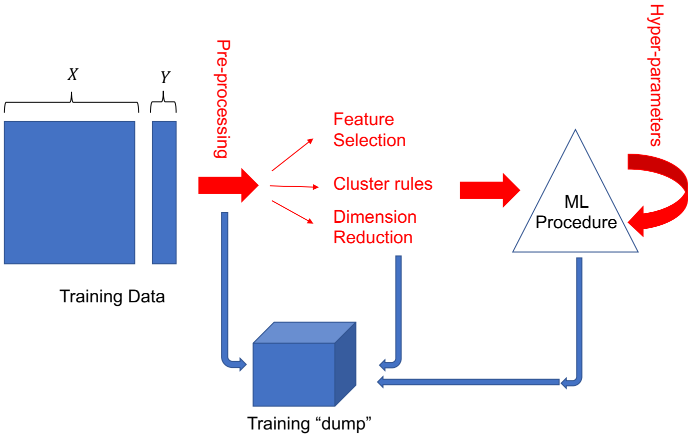




Goal

Let's classify documents about renaissance artists and the corresponding TMNT

Analysis flowchart: Reminder

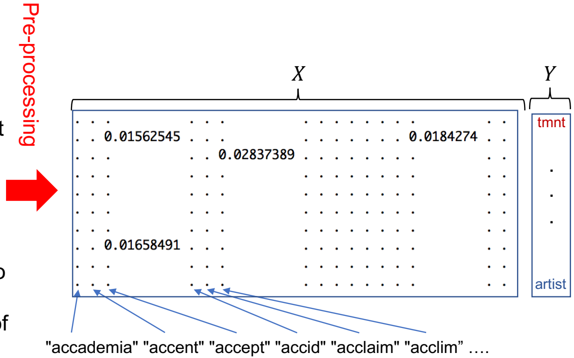


Preprocessing

"… highly skilled in Ninjutsu Olympic-level agility, speed and strength Keen strategist Master of stealth…"

.
.
.

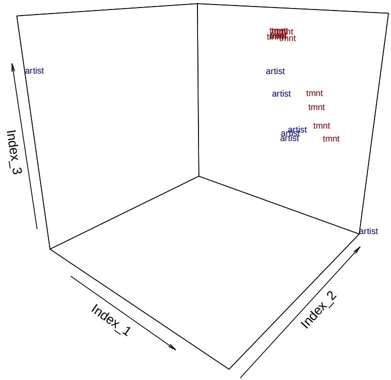
"… he was commissioned to finish the carving of the last small figures of the Shrine of St. Dominic, in the.…"



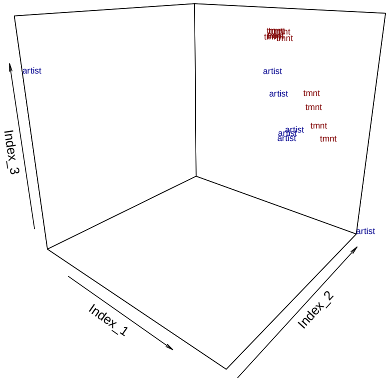
(Note: we wouldn't want to subtract column means from this data… why?)

$$X \in \mathbb{R}^{m \times n} \rightarrow m = 141 \text{ and } n = 5185$$

Dimension reduction



Dimension reduction



" ...Category View page ratings Rate this page What's this? Trustworthy Objective Complete Well-written I am highly knowledgeable about this topic (optional) Submit ratings ..."

ML procedure

- **Classification:**
  - (Sparse) Logistic Regression
  - Naive Bayes
  - Support Vector Machines (SVM)
  - k-Nearest Neighbors (KNN)
- **Regression:**
  - (Sparse) Linear Regression
  - Support Vector Regression
- **Both:**
  - Random Forest
  - Gradient Boosting Machines (GBM)
  - Neural Networks

ML procedure

- Classification:
  - (Sparse) Logistic Regression
  - Naive Bayes
  - Support Vector Machines (SVM)
  - k-Nearest Neighbors (KNN)
- Regression:
  - (Sparse) Linear Regression
  - Support Vector Regression
- Both:
  - Random Forest
  - Gradient Boosting Machines (GBM)
  - Neural Networks

ML procedure: Sparse logistic regression

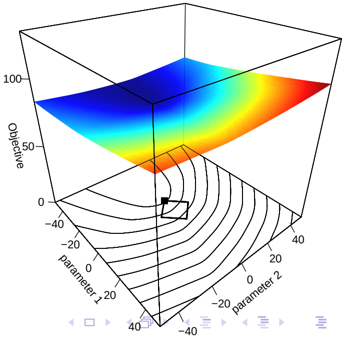
$$\pi(X) = \mathrm{Prob}(Y = \text{ artist } \mid X = [\text{accademia, accent, accept, accid,...}]^\top)$$

$$\log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta^\top X$$

$$\rightarrow \ell(\beta) = -\sum_{i=1}^{m}\left[Y_i X_i^\top \beta - \log\left(1 + e^{X_i^\top \beta}\right)\right] \quad \text{is the objective}$$

Constrained minimization $\longleftrightarrow$

$\rightarrow$ Use projected gradient descent



(1) $\hat\beta \overset{\text{update}}{=} \hat\beta - \eta\nabla\ell\big|_{\hat\beta}$

(2) $\hat\beta \overset{\text{set}}{=} \underset{\text{feasible } \beta}{\mathrm{argmin}} \; ||\hat\beta - \beta||$

ML procedure: Sparse logistic regression

Inference:

The magnitude/sign of $\hat{\beta}$ indicates which words affect the estimated probabilities the most

$\pi(X) = \mathrm{Prob}(Y = \text{ artist } | X = [\text{accademia, accent, accept, accid,...}]^\top)$

$\hat{\beta}$: positive, negative

Wrapping up

This approach has performed well on many of the problems I have worked on

Of course, nothing is perfect

It is important to keep on improving on what we have learned...

... just like in machine learning

Thank you!