Feature Engineering and Imputation

Rayid Ghani and Kit Rodolfa

Carnegie Mellon University

ML
MACHINE LEARNING DEPARTMENT

Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Reminders

This week:

- Tomorrow: Wednesday Deep Dive Session on Modeling and Validation Plans

Coming up next week:

- Monday: Project Update 3
- Tuesday: Weekly Feedback Form
- Thursday: Reading on Transductive Top-k

What We’ll Cover Today

- Feature Creation/Engineering
- Introducing Bias in Feature Development
- Dealing with Missing Data

Why do we care?

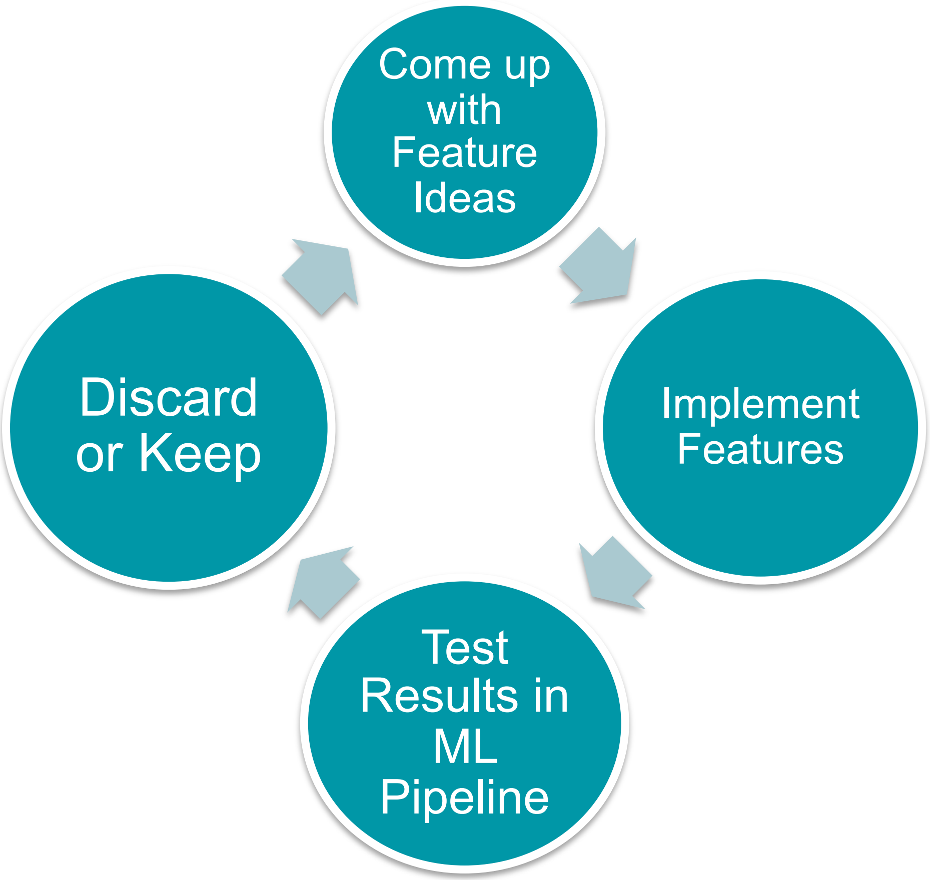
- Features are hints/rules of thumb you give your model
    - Encoding domain knowledge and context for the model to use
- Feature generation is a critical part of the machine learning modeling process, especially with structured data.
- Complexity in features may allow us to use less complex models that are faster to run, easier to understand and easier to maintain.

Practical Pointers

- When generating a feature, what did you know and **when** did you know it?
    - You can only create features from information available **before** the “training” date for a given row
- Domain/expert knowledge and prior research in the field can help a lot!

Feature development is an **iterative** process

Start simple: build a couple features (from each data source) you think are most important and expand from there



Bias in Feature Development: Mechanisms

- Is your feature directly measuring what you want it to or a proxy? Is it an equally good proxy across groups?

- Is measurement error correlated to group membership?

- How does predictiveness of your feature vary by group?

- Does missingness vary across groups?

Bias in Feature Development: Examples

- Inferring age/gender from name
- Creating "other" categories, e.g., multi-racial or non-binary gender
- How are race and ethnicity collected? Self-reported? Recorded by third party? Inferred from other data?
- Geocoding for distance or geographic features — how are homeless and more mobile populations handled?

Discussion Question

In the class project, what are two ways bias might be introduced in your feature development?

Feature Generation

- Categorical to Binary (Dummies)
- Features for missing values
- Discretization
- Date/Time Features
- Scaling/Normalizing
- Transformations
- **Aggregations (space, time, space and time)**
- **Relative (compared to the average…)**
- Interactions

Categorical to Binary

- One vs All (Dummy Variables)
- Groups
- Presence vs Absence

Discretization

- Equal width bins
- Equal size bins
- Entropy-based bins
- Domain-Specific bins to incorporate domain specific discontinuities
    - Age in general
    - Education/school data
    - High school data
    - Infant mortality

Feature Scaling

- Usually a good idea to scale features to have similar range: [-1,1] or [0,1] for example
  - Be careful with outliers

- Standardize/Normalize
  - Zero mean and unit variance

$$x_{new} = \frac{x - \mu}{\sigma}$$

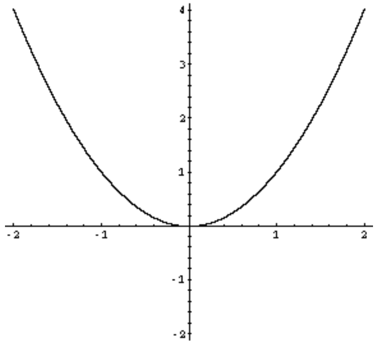
  - Sklearn.preprocessing.normalize

Is Scaling Important for…

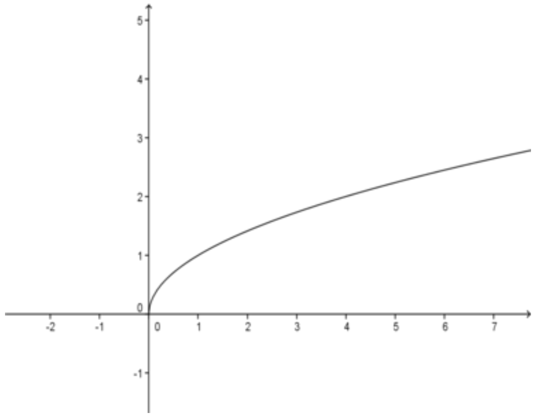
- Decision Trees?
- k-Nearest Neighbors?
- Random Forest?
- Logistic Regression?
- Neural Nets?

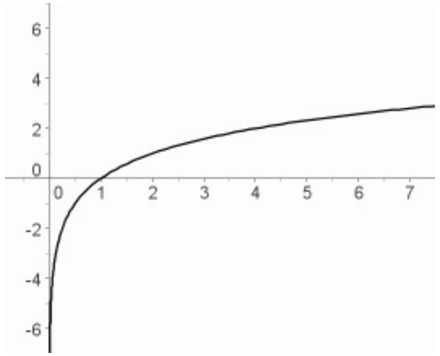
Feature Transformations

- Non-linear
  - Log (decreasing marginal utility)
  - (Square) Root
  - Squared



$y = x^2$





Aggregations

- Date differences (# of days since…)
- Aggregates over different time periods
  - min, max, avg, stdev
  - Avg spend in the past 3 months
- Relative aggregates
  - 1.5x avg spend
- Distances
- Aggregates over different distances
- Seasonality

Feature Interactions

- Generate features for combination of features
    - Age x gender

- Allows you to use linear models but still model non linear relationships

- Random Forests are one way of discovering useful interactions

Features are also model-dependent

- Linear models may need … ?
- Non-linear models may need …?

Missing Values

- Impute (Fill in) missing values based on why you think they may be missing and what you want the model to do with those missing values
    - Missing Completely at Random
    - Missing at Random
    - Missing Not at Random

- Typically, also add binary feature (dummy) for missing vs not missing in case “missingness” is predictive of the outcome

Imputing Missing Values: Some Options

- Nothing?
- Central Tendency: Mean / Median / Mode
- ML methods that handle missing data (e.g., xgboost)
- Others
    - Regression
    - k-Nearest Neighbor
    - Multiple Imputation

Imputing – Central Tendency

- Simple to calculate and computationally fast
- Often a reasonable starting point
- May be able to capture more nuance by using other, correlated data to help fill in missing values
- Under-represents variance/covariance of data

Imputing – ML Methods with Missing Data Handling

- Some models have built-in handling of missing data, such as xgboost (which decides which direction to send missing values at each split)

- May not be the best modeling method for your problem, don’t want to be locked into certain type of model

- Nevertheless, worth exploring performance of other imputation methods even when using these models as well

Imputing – Regression

- Make use of information in correlated features, more flexible than central tendency
- However, may not be flexible enough to capture complex relationships or interactions
- May be somewhat more computationally expensive
- Generally will still underestimate variation in data

Imputing – k-Nearest Neighbor

- More flexible option, capture more complexity in relationships in data
- Difficult to choose appropriate distance metric, value of k
- More computationally expensive than other methods of imputation
- Requires entire training set to calculate imputed values for new examples

Imputing – Multiple Imputation

- Create multiple “complete” datasets with different values using different regression models

- Helps analyze sensitivity to handling of missing values

- Much more computationally expensive, both for imputation and downstream modeling

- Can provide better representation of variability in data

Missing Value Tips

- Do not remove rows or columns with missing values (unless there is a really really really good reason)

- Missingness can be a useful predictor: create a flag even if you impute a value

- Data can be missing for different reasons and missingness for each row/column/cell may need to be handled differently

- Only use data from the past for imputation

How can imputation introduce bias in your models?

Reminders

This week:

- Tomorrow: Wednesday Deep Dive Session on Modeling and Validation Plans

Coming up next week:

- Monday: Project Update 3
- Tuesday: Weekly Feedback Form
- Thursday: Reading on Transductive Top-k