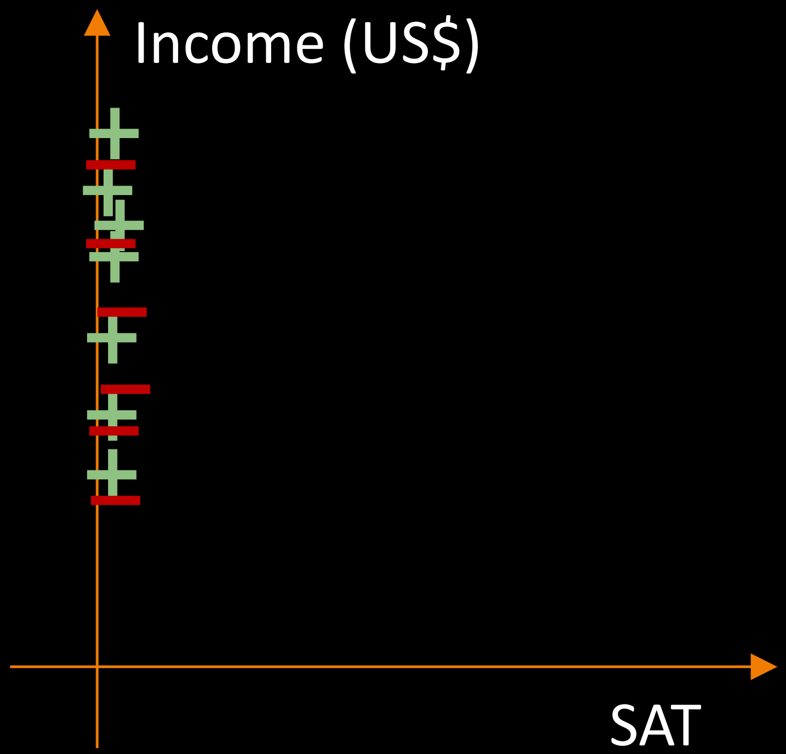
Normalizing Features

CS4780/5780 – Introduction to Machine Learning

Thorsten Joachims
Cornell University

Why Normalize?

- K-NN with Euclidian Distance?
  - SAT barely affects distance
- Linear SVM
  - $\vec{w} = (w_{SAT}, w_{Income})$ needs huge value for w_{GPA} → small margin
- Neural Network
  - SGD update makes big steps on w_{Income} and small steps on w_{SAT}
- Etc.



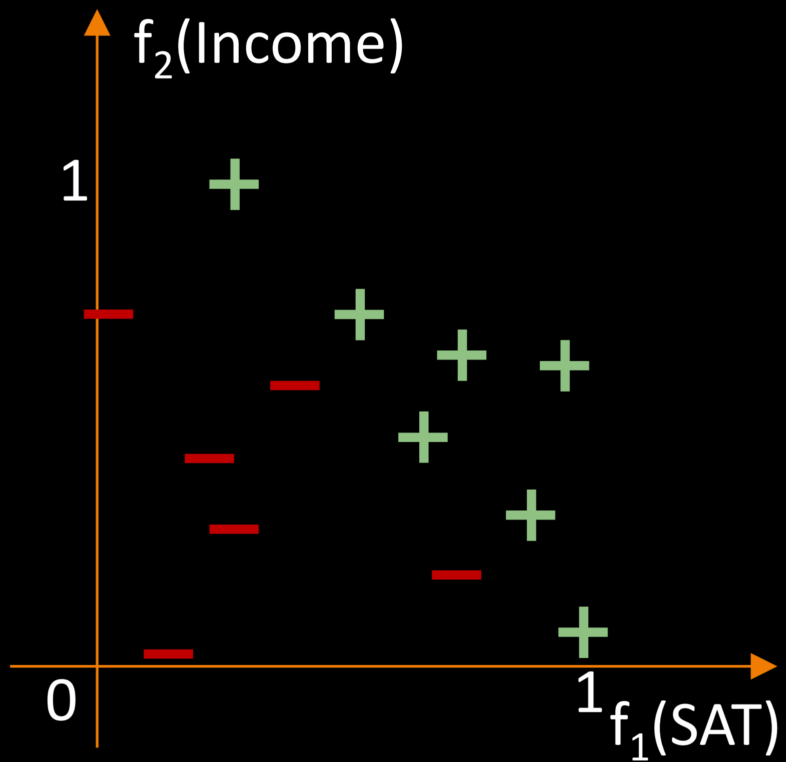
Y=Attend College?

Feature Scaling

- Idea: Transform features $X_1 \dots X_N$ so that they all lie on the same scale.
- Process:
  1. Analyze the training set and determine transformation $f_j\colon X_j \rightarrow \bar{X}_j$ for each feature X_j.
     $\rightarrow f(f_1, \dots, f_N)$
  2. Apply transformation $f(\vec{x}_i)$ to all training examples $\vec{x}_i \in S_{Train} \rightarrow \bar{S}_{Train}$
  3. Apply transformation $f(\vec{x}_i)$ to all test examples $\vec{x}_i \in S_{Test} \rightarrow \bar{S}_{Test}$

Min/Max Scaling

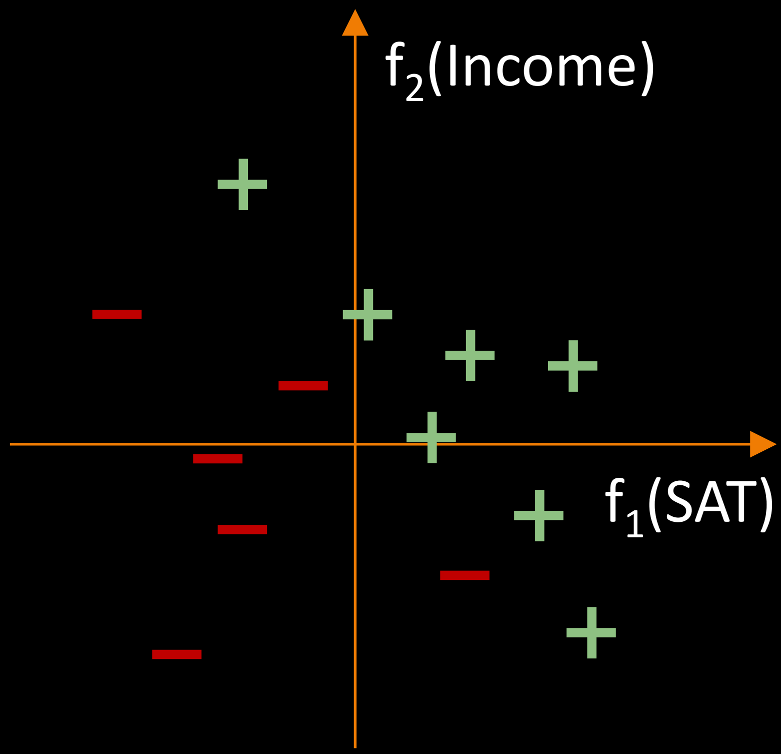
- Idea: Scale features so that the minimum and the maximum value of all features become equal.
  - For each feature X_j
    - Find minimum $\min(X_j)$ and maximum $\max(X_j)$ in training sample.
    - $f_j(x_j) = \frac{x_j - \min_j}{\max_j - \min_j}$



Y=Attend College?

Standardization

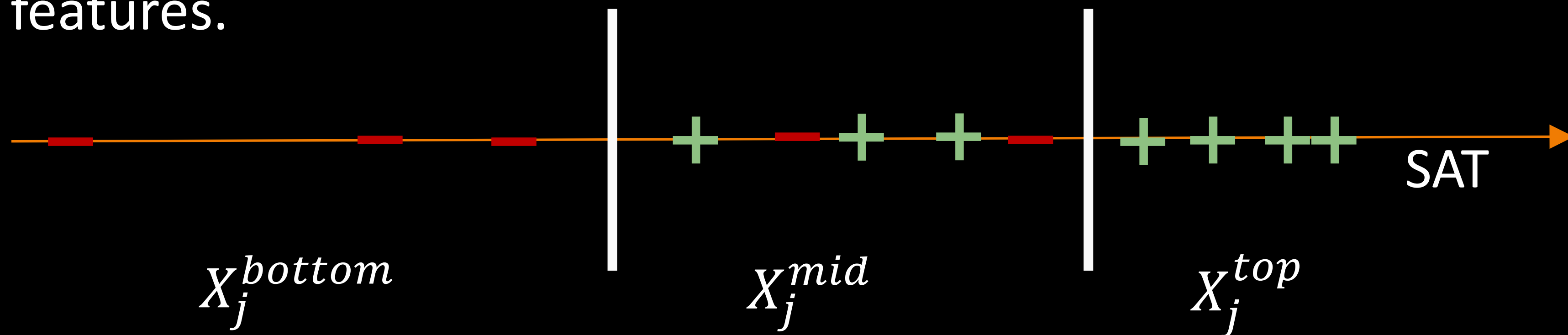
- Idea: Scale features so that the mean and standard deviation of all features become equal.
  - For each feature X_j
    - Estimate mean via $\text{average}(X_j)$ and standard deviation via $\sqrt{\text{average}(X_j^2) - \text{average}(X_j)^2}$ in training sample.
    - $f_j(x_j) = \frac{x_j - \text{mean}_j}{\text{stddev}_j}$



Y=Attend College?

Percentile Binning

- Idea: Transform numeric feature X_j into multiple binary features.



- More general:
  - One binary feature for each percentile.
    - Student has SAT above 1500
    - Student has SAT above 1400
    - Student has SAT above 1000
    - Student has SAT above 500

→ Multiple binary features can be “1” for each student.

SAT of 1100 → (0,0,1,1)

SAT of 300 → (0,0,0,0)

Missing Values

- Problem: For some feature X_j, the feature value may be missing for some examples.
  - Not taken SAT → $X_{SAT} = NULL$
- Simplest approach
  - Impute value for X_j (e.g. value zero)
  - Indicate that value is imputed via new binary feature

$$X_j^{Missing} = \begin{cases} 1 & X_j = NULL \\ 0 & otherwise \end{cases}$$

Summary

- Typically need to transform features
  - Use same transformation for train and test examples
- Feature scaling relates to how the learning algorithm interprets them
- Indicate and impute missing features