Loss Functions for Preference Levels: Regression with Discrete Ordered Labels

Jason Rennie
Massachusetts Institute of Technology

Nati Srebro
University of Toronto
Supervised Learning Setting (Regression)

• Given a labeled training set:

\[
\begin{align*}
  x_1 & \quad y_1 \\
  x_2 & \quad y_2 \\
  \vdots \\
  x_n & \quad y_n
\end{align*}
\]

objects, e.g. movies, options, etc. described as a feature vector

• Learn a mapping

\[ f(x) \mapsto y \]

in order to predict labels on future data:

\[
\text{?}
\]
Target Labels

• Common types of target labels:
  – Binary (positive/negative; 😞 😊)
  – Multiclass (discrete, unordered categories)
  – Real valued

• Discrete ordinal labels
  ★ ★★★ ★★★★★ ★★★★★★★★
  😞 ☹ ☻
  “undesirable”, “indifferent”, “preferred”
Background: Binary Regression

- Labeled training set \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\)
- Learn \(z(x) = w'x + w_0\)
  such that \(z(x) > 0\) when \(y = +1\),
  and \(z(x) < 0\) when \(y = -1\)

minimizing \(\sum_i \text{loss}(z(x_i); y_i)\)

\[
\text{loss}(z; y) = \begin{cases} 
0 & \text{if } yz > 0 \\
1 & \text{otherwise}
\end{cases}
\]
Background: Binary Regression

- Labeled training set \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\)
- Learn minimizing

\[ z(x) = w'x + w_0 \]
\[ \sum_i \text{loss}(z(x_i);y_i) + \lambda |w|^2 \]

\[ \text{loss}(z;y) = \begin{cases} 0 & \text{if } yz > 1 \\ 1 & \text{otherwise} \end{cases} \]

Focus on linear regression as an example. Same ideas apply to any other family of predictors.

SVM
logistic regression
L2 SVM
Discrete Ordinal Labels

• Instead of \( y = -1 \) or \( +1 \), we have \( y = 1, 2, 3, \ldots, k \)

• Treat as \( k \) multiple unrelated classes, learn separate classifier for each value?

• Treat as a real valued objective, minimize, e.g. sum-squared error?
Threshold based approach

\[ z(x) = w^T x \]

\[ y = +1 \]

\[ \text{loss}(z(x); y) \]
Threshold based approach

\[
\text{Immediate-threshold loss} \\
\text{all-threshold loss}
\]

\[\text{loss}(z(x); y)\]

\[z(x) = w'x\]

[Shashua Levin 03]
Threshold based approach

- All-threshold loss is a bound on the absolute rank-difference

- For both constructions:
  - can use any penalty function (e.g. logistic) instead of hinge
  - learn per-user $\theta$’s (different users use ratings differently)
Results on MovieLens Data

Least squares: 1.33

Mean Absolute Error

Least squares: 0.76

Zero-One Error

All-Threshold vs others significant at p<10^{-16}

All-Threshold vs others barely significant at p<0.14

[Bar charts showing performance metrics for different classification thresholds and models, including Least squares, Truncated Square Error, (Smoothed) Hinge, and Logistic.]
Beyond Linear Regression

- Same constructions can be used whenever a loss function is needed:
  - Kernel methods (SVMs)
  - Collaborative prediction (matrix completion)

[Srebro Rennie Jaakkola NIPS’04]
[Rennie Srebro ICML’05]
Other Loss Functions

- Generalization to the logistic motivated by probabilistic generative model (see paper)
- Similar generative model with additive Gaussian “noise” [Chu Ghahramani 2004]

Alternative approach:
- Map ordinal labels to “<“ relationships [Herbrich et al 2000]
  – quadratic number of relationships
Summary

• Studied different constructions for loss-functions for discrete ordinal labels

• All-threshold construction best, much better then treating as multiclass or using squared error

• Can be used whenever a (scale sensitive) loss function is needed

www.cs.toronto.edu/~nati