Predicting Regression Probability Distributions

with Imperfect Data

through Optimal Transformations

Jerome H. Friedman

Stanford University
MACHINE LEARNING / REGRESSION

\[ y = F(x, z) \]

\( y \) = outcome variable

\( x = (x_1 \ldots, x_p) \) observed predictor variables

\( z = (z_1, z_2, \ldots) \) other variables

Goal: estimate \( E[y \mid x] \) given data \( \{y_i, x_i\}_{i=1}^N \)
STATISTICAL MODEL

\[ y = f(x) + s(x) \cdot \epsilon \]

\[ f(x) = E[y \mid x] \quad \text{location function} \]

\[ s(x) > 0 \quad \text{scale function} \]

\[ \epsilon = \text{random variable, } E[\epsilon \mid x] = 0 \]

Prediction: \( \hat{y} = f(x) \)

\[ s(x) \cdot \epsilon = \text{"irreducible error" (unavoidable)} \]
REDUCIBLE ERROR

\[ r(x) = E | f(x) - \hat{f}(x) | \]

\( f(x) \) = optimal location (target) function

\( \hat{f}(x) \) = estimate based on training data & ML method

ML goal: methods to reduce \( r(x) \)

Statistics goal: methods to estimate \( r(x) \)

Prediction error \( (y) = \text{Reducible} + \text{Irreducible} \)

Usually: Irreducible \( s(x) \gg\) Reducible \( r(x) \)
USUAL ASSUMPTIONS

\[ s(x) = s = \text{constant} \text{ (homoscedasticity)} \]

\[ \epsilon \sim N(0, 1) \text{ (normality)} \]

Neither very likely

Tukey:

"small residuals \sim\text{ normal, larger have heavier tails.}"
LOGISTIC DISTRIBUTION

\[ \epsilon | x = (y - f(x)) / s(x) \]

\[ \tilde{p}(\epsilon) = \frac{e^{-\epsilon}}{s(1+e^{-\epsilon})^2} \]

small \( |\epsilon| \sim \text{normal} \), large \( |\epsilon| \sim \text{exponential} \)
Prediction: \( \hat{y} = \hat{f}(x) \)

\[
\hat{f}(x) = \arg \min_{f \in F} \sum_{i=1}^{N} [\varepsilon_i + 2 \log(1 + e^{-\varepsilon_i})]
\]

\( \varepsilon_i = (y_i - f(x_i))/s(x_i) \)

minimized at \( f(x_i) = y_i \) indep \( s(x_i) \)

\( 1/s(x_i) \sim \) “weight” for obs \( i \)

controls relative influence of \( i \) to fit
Using incorrect $s(x)$ to estimate $f(x)$

increases variance, not bias

assume $s(x) =$ constant usually not too bad
ESTIMATE $\hat{s}(x)$

(1) Improve $\hat{f}(x)$ in high variance settings.

(2) Important inferential statistic:

   (a) prediction interval $\sim$ accuracy of $\hat{y}$-prediction:

   logistic: $\text{IQR}[y \mid x] = 2 s(x) / \log(3)$

   (b) can affect decision
CENSORING

Data: \( \{y_i, x_i\}_1^N \rightarrow \{a_i, b_i, x_i\}_1^N \)

\[ a_i \leq y_i \leq b_i \]

\[ a_i = b_i = y_i \Rightarrow \text{y-value known} \]

\[ a_i = -\infty \Rightarrow \text{censored below } b_i \]

\[ b_i = \infty \Rightarrow \text{censored above } a_i \]

Otherwise: interval censored \([a_i, b_i]\)
LIKELIHOOD

\[ \Pr(a \leq y \leq b) = \frac{1}{1+e^{-(b-f)/s}} - \frac{1}{1+e^{-(a-f)/s}} \]

Depends strongly on both \( f \) and \( s \)

Need to estimate both \( f(x) \) and \( s(x) \)
EXERCISE

\[ (\hat{f}(x), \hat{s}(x)) = \arg \min_{f, s \in F} \sum_{i=1}^{N} L(a_i, b_i, f(x_i), s(x_i)) \]

\[ L(a, b, f(x), s(x)) = \log \left[ \frac{1}{1 + \exp((f(x) - a)/s(x))} - \frac{1}{1 + \exp((f(x) - b)/s(x))} \right] \]
GRADIENT BOOSTED TREE ENSEMBLES


\[
\hat{f}(x) = \sum_{k=1}^{K_f} T^{(f)}_k(x)
\]

\[
\log(s(x)) = \sum_{k=1}^{K_s} T^{(s)}_k(x)
\]

\[
T^*_k(x) = \text{CART-tree}(x)
\]
ITERATIVE GRADIENT BOOSTING

Start: \( \hat{s}(x) = \text{constant} \)

Loop 

\[
\hat{f}(x) = \text{tree-boost} \left[ f(x) \mid \hat{s}(x) \right]
\]

\[
\log(\hat{s}(x)) = \text{tree-boost} \left[ \log(s(x)) \mid \hat{f}(x) \right]
\]

Until no change
OPTIMAL TRANSFORMATIONS

\[ g(y) = f(x) + s(x) \cdot \varepsilon \]

\[ g(y) = \text{unknown monotonic function} \]

\[ [\hat{g}(y), \hat{f}(x), \hat{s}(x)] = \arg \min_{g, f, s} \sum_{i=1}^{N} L[g(y_i), f(x_i), s(x_i)] \]

\[ \frac{1}{N} \sum_{i=1}^{N} CDF(\hat{g}(y), \hat{f}(x_i), \hat{s}(x_i)) = CDF_y(y) \]

\[ CDF(g(y), f(x), s(x)) = \frac{1}{1+\exp((f(x)-g(y))/s(x))} \]
ASYMMETRY

Asymmetric Logistic

\[
p(y)
\]

\[
y
\]
ASYMMETRIC LOGISTIC DISTRIBUTION

\[
p(z \mid f, s_l, s_u) = \frac{2}{s_l + s_u} \left[ \frac{I(z \leq f) \exp((f - z)/s_l)}{(1 + \exp((f - z)/s_l))^2} + \frac{I(z > f) \exp((f - z)/s_u)}{(1 + \exp((f - z)/s_u))^2} \right]
\]

\[
g(y) = f(x) + \eta, \quad \begin{cases} 
\eta = -s_l(x) \cdot |\varepsilon|, & \text{prob} = s_l(x)/(s_l(x) + s_u(x)) \\
\eta = +s_u(x) \cdot |\varepsilon|, & \text{prob} = s_u(x)/(s_l(x) + s_u(x)) 
\end{cases}
\]

\[
f = \text{mode}, \ s_l = \text{lower scale}, \ s_u = \text{upper scale}
\]
\[
\varepsilon \sim L(0, 1)
\]
ASYMMETRIC GRADIENT BOOSTING

Start: \( \hat{s}_l(x) = \hat{s}_u(x) = \text{constant} \)

Loop {

\[ \hat{f}(x) = \text{tree-boost } f(x) \text{ given } \hat{s}_l(x) \& \hat{s}_u(x) \]

\[ \log(\hat{s}_l(x)) = \text{tree-boost } s_l(x) \text{ given } \hat{f}(x) \& \hat{s}_u(x) \]

\[ \log(\hat{s}_u(x)) = \text{tree-boost } s_u(x) \text{ given } \hat{f}(x) \& \hat{s}_l(x) \]

} Until change < threshold.
OPTIMAL TRANSFORMATIONS

\[ g(y) \mid x \sim \text{Logistic}(\hat{f}(x), \hat{s}_l(x), \hat{s}_u(x)) \]

\[
[\hat{g}(y), \hat{f}(x), \hat{s}_l(x), \hat{s}_u(x)]
\]

\[ = \arg\min_{g,(f,s_l,s_u) \in F} \sum_{i=1}^{N} L[g(y_i), f(x_i), s_l(x_i), s_u(x_i)] \]

\[
\frac{1}{N} \sum_{i=1}^{N} CDF(\hat{g}(y_i), \hat{f}(x_i), \hat{s}_l(x_i), s_u(x_i)) = CDF_y(y)
\]

\[ CDF(g(y), f(x), s(x)) = \text{messy closed form} \]
DIAGNOSTICS

For $x_i \in S$:

Model: $\Pr(\hat{g}(y) < z) = \frac{1}{|S|} \sum_{i \in S} CDF(z, \hat{f}(x_i), \hat{s}_i(x_i), \hat{s}_u(x_i))$

Empirical: $\hat{\Pr}(\hat{g}(y) < z) = \frac{1}{|S|} \sum_{i \in S} I(\hat{g}(y_i) \leq z)$

Symmetric: $(g(y_i) - f(x_i))/s(x_i) \sim L(0, 1)$

Asymmetric: $I(-) (\hat{g}(y_i) - \hat{f}(x_i))/\hat{s}_i(x_i)$

$+ I(+) (\hat{g}(y_i) - \hat{f}(x_i))/\hat{s}_u(x_i) \sim L(0, 1)$
DATA SUBSETS $S$

$$S = \{i \mid r < \hat{f}(x_i) \leq t \& \ u < \hat{s}(x_i) \leq v\}.$$  \hspace{1cm} (1)

$$S = \{i \mid r < \hat{f}(x_i) \leq t \& \ u < \sqrt{\hat{s}_l(x_i)\hat{s}_u(x_i)} \leq v\}.$$  \hspace{1cm} (2)

$$S = \{i \mid r < \hat{f}(x_i) \leq t \& \ u < \hat{s}_l(x_i) \leq v\}.$$  \hspace{1cm} (3)

$$S = \{i \mid r < \hat{f}(x_i) \leq t \& \ u < \hat{s}_{ul}(x_i) \leq v\}.$$  \hspace{1cm} (4)
QUESTIONNAIRE DATA ($N = 8856$)

Bay Area Shopping Mall Customers

AGE ∈

1  2  3  4  5  6  7
17 & under 18–24 25–34 35–44 45–54 55–64 65 & older

x =

1 Occupation 8 Dual Incomes
2 Type of home 9 Persons in household
3 Gender 10 Persons in household under 18
4 Martial status 11 Householder status
5 Education 12 Ethnic classification
6 Annual income 13 Language
7 Lived in Bay Area
Q–Q plots of empirical versus predicted quantiles for untransformed Omnireg solution.
MASHABLE ONLINE NEWS POPULARITY

(Irvine Repository)

\[ N = 39644 \text{ news articles} \]

25000 train, 7322 mod. sel., 7322 test

\[ y = \text{popularity (number of shares on social media)} \]

\[ x = 59 \text{ numerical attributes} \]
(1/2) – Million Song Dataset

Song Recordings

(Irvine Repository)

\[ N(train) = 463715 \text{ song recordings} \]

50000 train, 10000 mod. selection

\[ N(test) = 51630 \]

\( y = \) year released (1922 – 2011)

\( x = 89 \) acustic measurements
LESSONS

1. Don’t confuse reducible and irreducible error.

2. Tukey was right (normality is rare)

3. Even approximate homoscedasticity is rare

4. Even approximate symmetry is rare

5. Optimal (nonobvious) transformations can help
Slides: