PRESTO! Predicting Rare Events by Shrinking Towards Proportional Odds

Gregory Faletto

Ph.D. Candidate, University of Southern California Marshall School of Business

Department of Data Sciences and Operations

December 2nd, 2022

Joint work with Jacob Bien

Outline

Estimating Probabilities of Rare Events is Important and Hard

Classifiers struggle in the *class imbalance* setting where one class is very rare.

Unfortunately, accurate estimates of the probabilities of rare events are often very important.

> **Online marketing:** clicking on ads and making a purchase is very rare. Important to accurately estimate the probability of a user making a purchase: click ads are sold by auction. Probability of purchase is needed to bid effectively.
>
> **Health and medicine:** rare diseases can be expensive to treat, the resources to deliver effective treatment are scarce, and it can be very stressful to believe you may have a rare disease. Accurately estimating the probability of having a rare disease is crucial to treat patients effectively.

Brief Recap: Bayes Decision Boundary

Setting: binary classification with $y \in \{1,2\}$

Definition: The Bayes decision boundary is the set of covariates $\boldsymbol{x} \in \mathbb{R}^p$ satisfying $\mathbb{P}(y = 1 \mid \boldsymbol{x}) = \mathbb{P}(y = 2 \mid \boldsymbol{x}) = 0.5$.

On one side of the decision boundary, class 1 is more likely; on the other, class 2 is more likely. (See Section 2.2 of James et. al 2021).

Brief Recap: Bayes Decision Boundary

Setting: binary classification with $y \in \{1,2\}$

Definition: The Bayes decision boundary is the set of covariates $\boldsymbol{x} \in \mathbb{R}^p$ satisfying $\mathbb{P}(y = 1 \mid \boldsymbol{x}) = \mathbb{P}(y = 2 \mid \boldsymbol{x}) = 0.5$.

On one side of the decision boundary, class 1 is more likely; on the other, class 2 is more likely. (See Section 2.2 of James et. al 2021).

Example: the logistic regression model for $\mathbb{P}(y = 1 \mid \boldsymbol{x})$ is

$$\log\left(\frac{\mathbb{P}\left(y = 1 \mid \boldsymbol{x}\right)}{\mathbb{P}\left(y = 2 \mid \boldsymbol{x}\right)}\right) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}\,.$$

for an intercept $\alpha \in \mathbb{R}$ and coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$. Bayes decision boundary: $\left\{\boldsymbol{x} : \alpha + \boldsymbol{\beta}^\top \boldsymbol{x} = 0\right\}$.

So estimating $(\alpha, \boldsymbol{\beta})$ also estimates the Bayes decision boundary. **Estimating the Bayes decision boundary well is more or less equivalent to estimating probabilities well.**

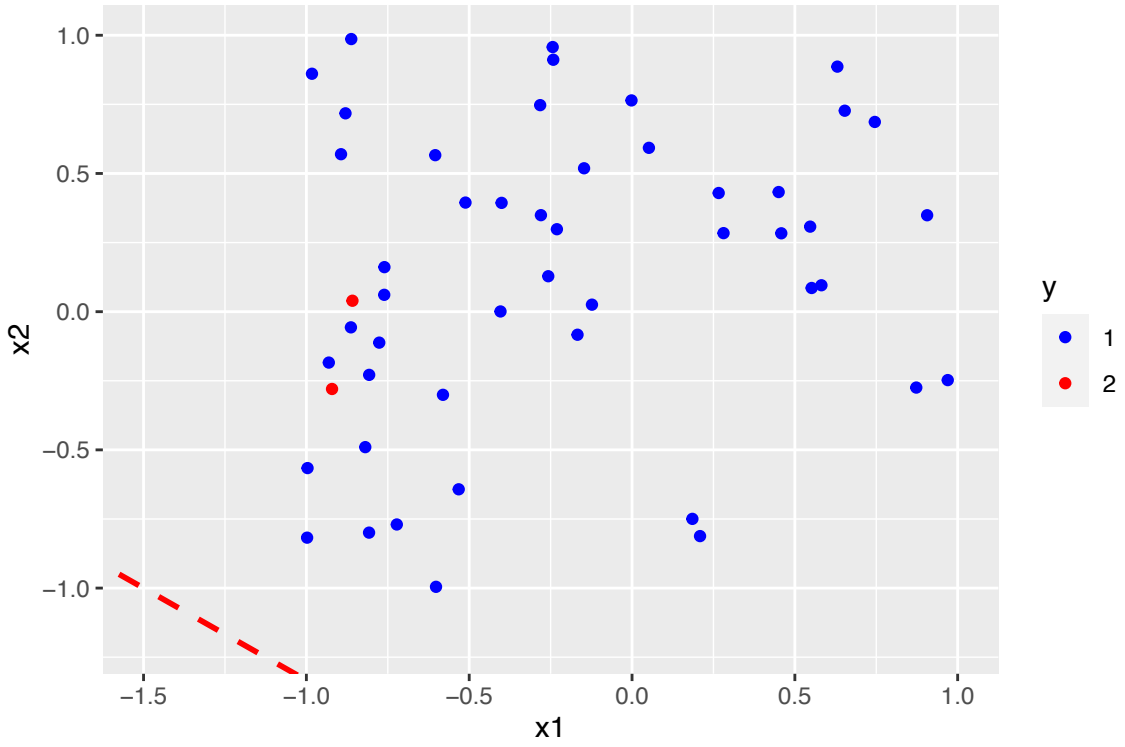
Conceptualizing the Problem

We draw 50 random $\boldsymbol{x} \in \mathbb{R}^2$. Each $x_{ij} \sim \text{Uniform}(-1,1)$ iid.

Binary label y generated from a logistic regression model.

Dashed line: Bayes decision boundary

Class 2 is rare $\iff$ we don't observe many points near the decision boundary. Estimating the decision boundary, and $\mathbb{P}(y = 2 \mid \boldsymbol{x})$, is very difficult.



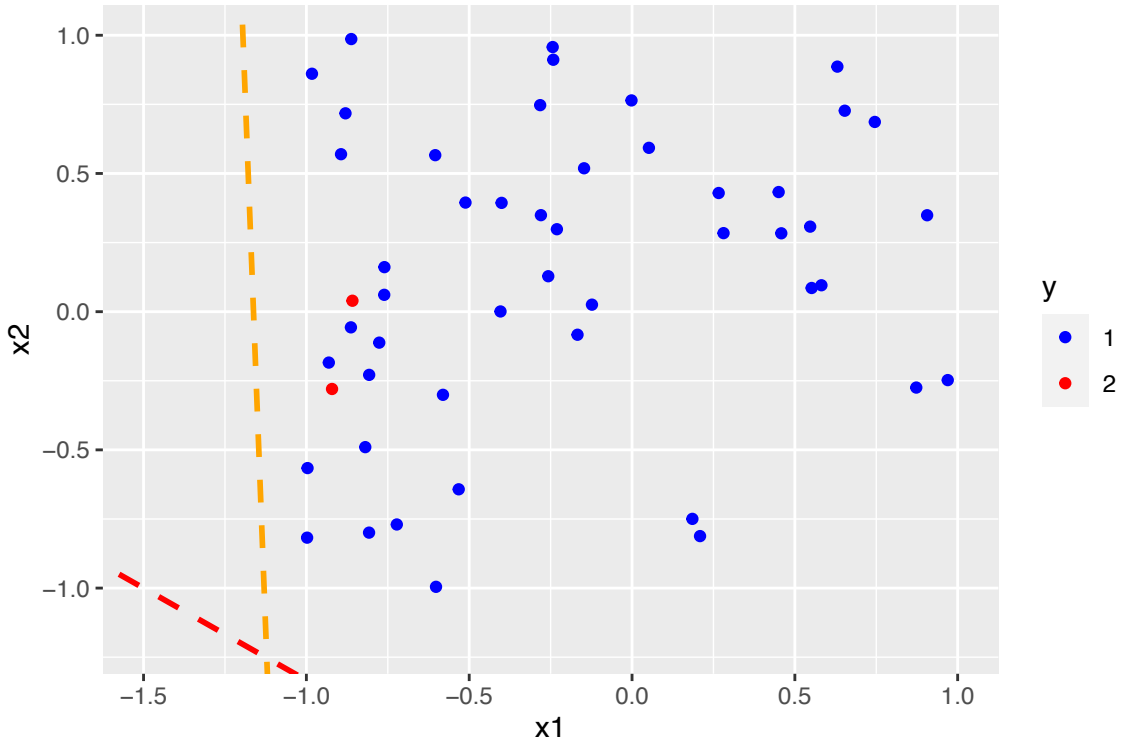
Conceptualizing the Problem

We draw 50 random $\boldsymbol{x} \in \mathbb{R}^2$. Each $x_{ij} \sim \text{Uniform}(-1,1)$ iid.

Binary label y generated from a logistic regression model.

Dashed line: Bayes decision boundary

Class 2 is rare $\iff$ we don't observe many points near the decision boundary. Estimating the decision boundary, and $\mathbb{P}(y = 2 \mid \boldsymbol{x})$, is very difficult.



The estimated decision boundary (orange) is very bad $\implies$ very bad rare probability estimates. (Even in large samples!)

A Possible Solution: Leveraging Data From More Common Outcomes

Sometimes there are more common intermediate (**ordered**) outcomes.

> **Online marketing:** clicking on ads without making a purchase is much more common than clicking on ads and making a purchase.
>
> **Health and medicine:** sometimes there are intermediate outcomes between having a disease and being completely healthy (normal blood glucose → prediabetes → diabetes, normal blood pressure → prehypertension → hypertension, etc.)

Estimating the decision boundaries between these more common outcomes is easier (less class imbalance).

A Possible Solution: Leveraging Data From More Common Outcomes

Sometimes there are more common intermediate (**ordered**) outcomes.

> **Online marketing:** clicking on ads without making a purchase is much more common than clicking on ads and making a purchase.
>
> **Health and medicine:** sometimes there are intermediate outcomes between having a disease and being completely healthy (normal blood glucose → prediabetes → diabetes, normal blood pressure → prehypertension → hypertension, etc.)

Estimating the decision boundaries between these more common outcomes is easier (less class imbalance).

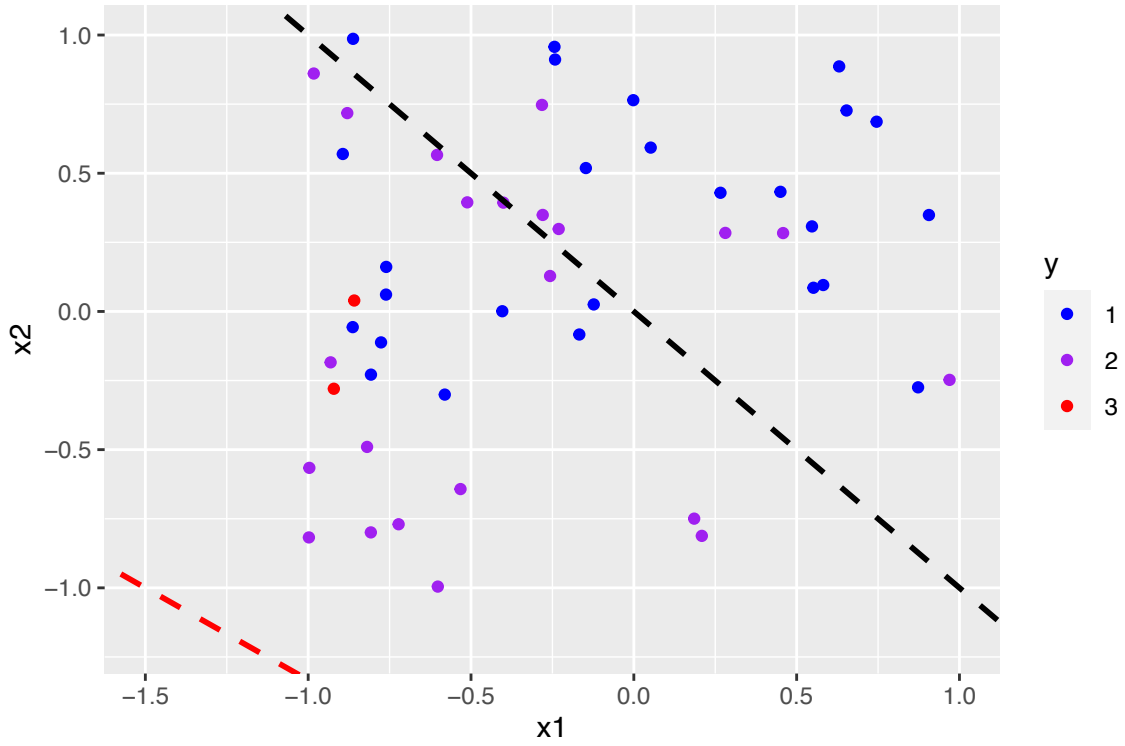
Idea: maybe the factors that are predictive of common outcomes can be used to improve predictions of rare outcomes.

A Possible Solution: Leveraging Data From More Common Outcomes

Same data as before, same red decision boundary as before. Class 2 is an intermediate outcome between classes 1 and 3. (In previous slide, classes 1 and 2 were combined.)

Black dashed line: Bayes decision boundary between classes 1 and 2.

Abundant data near decision boundary between classes 1 and 2 $\Longrightarrow$ easier to learn.



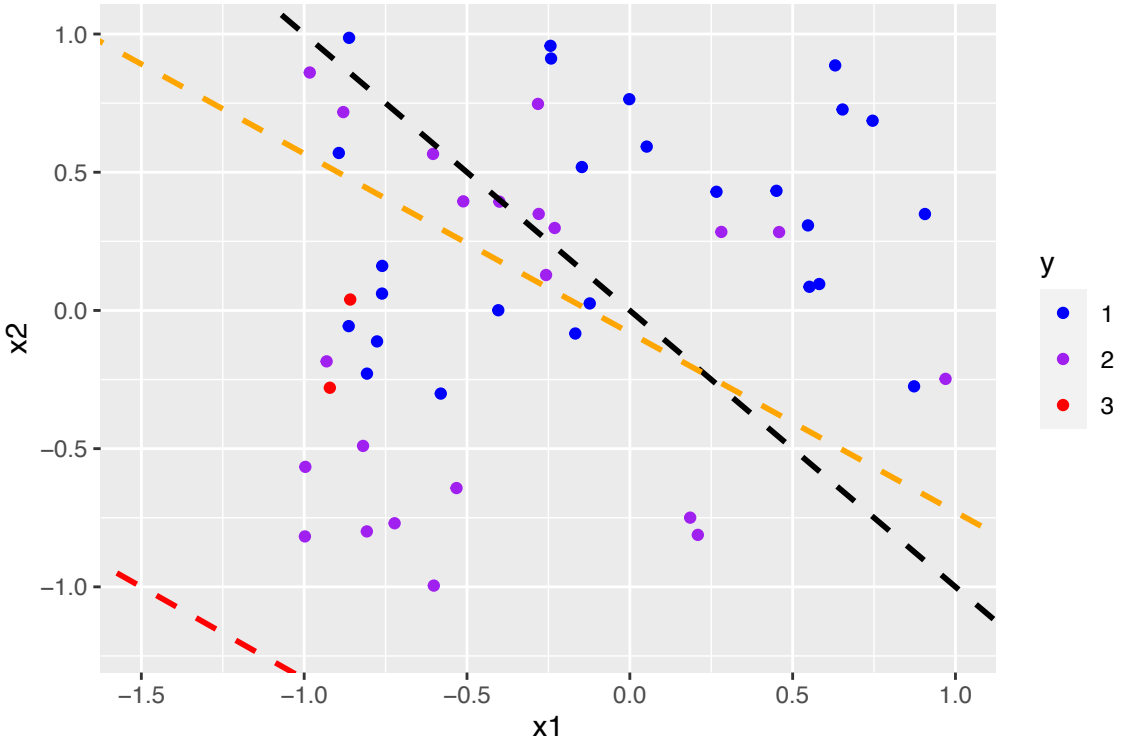
A Possible Solution: Leveraging Data From More Common Outcomes

Same data as before, same red decision boundary as before. Class 2 is an intermediate outcome between classes 1 and 3. (In previous slide, classes 1 and 2 were combined.)

Black dashed line: Bayes decision boundary between classes 1 and 2.

Abundant data near decision boundary between classes 1 and 2 $\Longrightarrow$ easier to learn.

The estimated decision boundary between classes 1 and 2 (orange) does seem to fit much better.



Maybe we can leverage our more precise estimate of the decision boundary between classes 1 and 2 to better estimate the decision boundary between classes 2 and 3 $\Longrightarrow$ better estimated probabilities of lying in class 3.

A Classical Model Can Do This: The Proportional Odds Model

P*roportional odds model* or *ordered logit model* (McCullagh, 1980) for ordinal outcomes $k \in \{1, \ldots, K\}$ assumes that the decision boundaries between successive categories $y \in \{1, \ldots, K\}$ all have the same $\boldsymbol{\beta}$ vector separated by $K - 1$ intercepts:

$$\log\left(\frac{\mathbb{P}\left(y \leq k \mid \boldsymbol{x}\right)}{\mathbb{P}\left(y > k \mid \boldsymbol{x}\right)}\right) = \alpha_k + \boldsymbol{\beta}^{\top}\boldsymbol{x}, \qquad k \in \{1, \ldots, K-1\}\,.$$

Equivalent: true model for the random variable $\mathbf{1}\{y > k\} \mid \boldsymbol{x}$ is logistic regression for all $k \in \{1, \ldots, K-1\}$ with all $\boldsymbol{\beta}$ restricted to be equal.

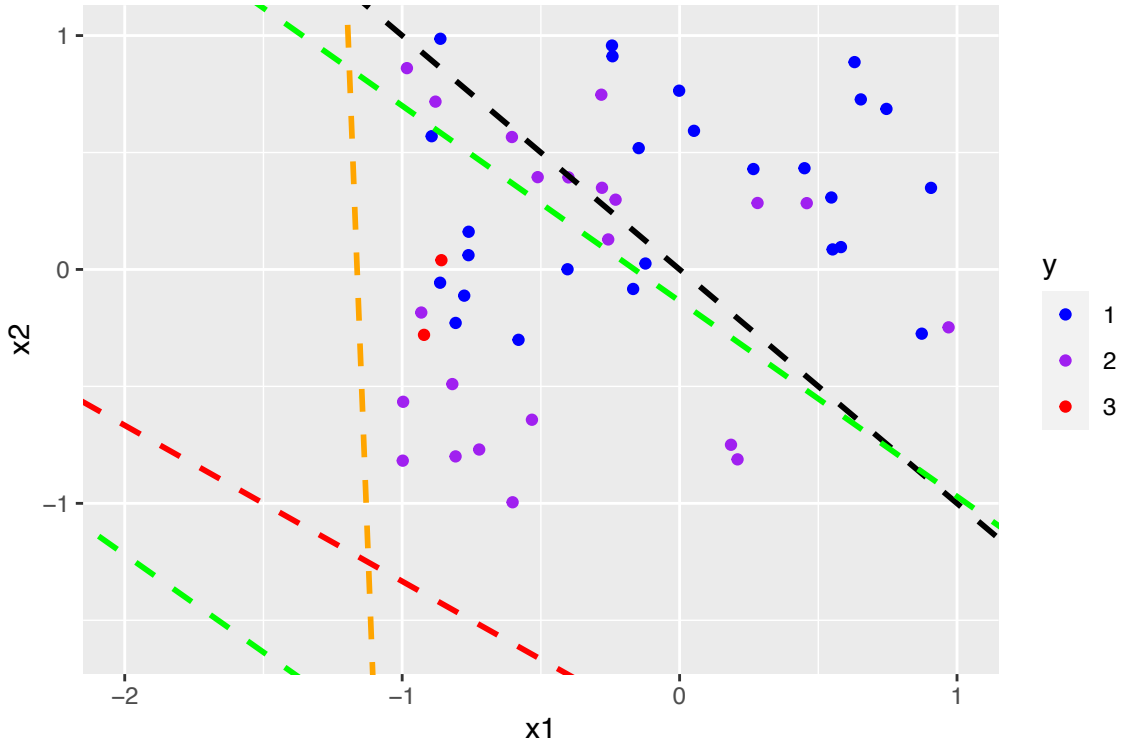
Only have to estimate one unique parameter for each of the $K - 1$ decision boundaries: α_k. (Assumption of equal $\boldsymbol{\beta}$ means we don't have to learn much from scarce data in rare categories.)

The Proportional Odds Model Leverages Data From More Common Outcomes

Green dashed lines: estimated decision boundaries from proportional odds model. (Orange decision boundary is from logistic regression on rare class, same as before.)

Proportional odds model does seem to yield a much better estimate of the rare decision boundary!

Also yields better estimates of $\mathbb{P}(y = 3 \mid \boldsymbol{x})$.

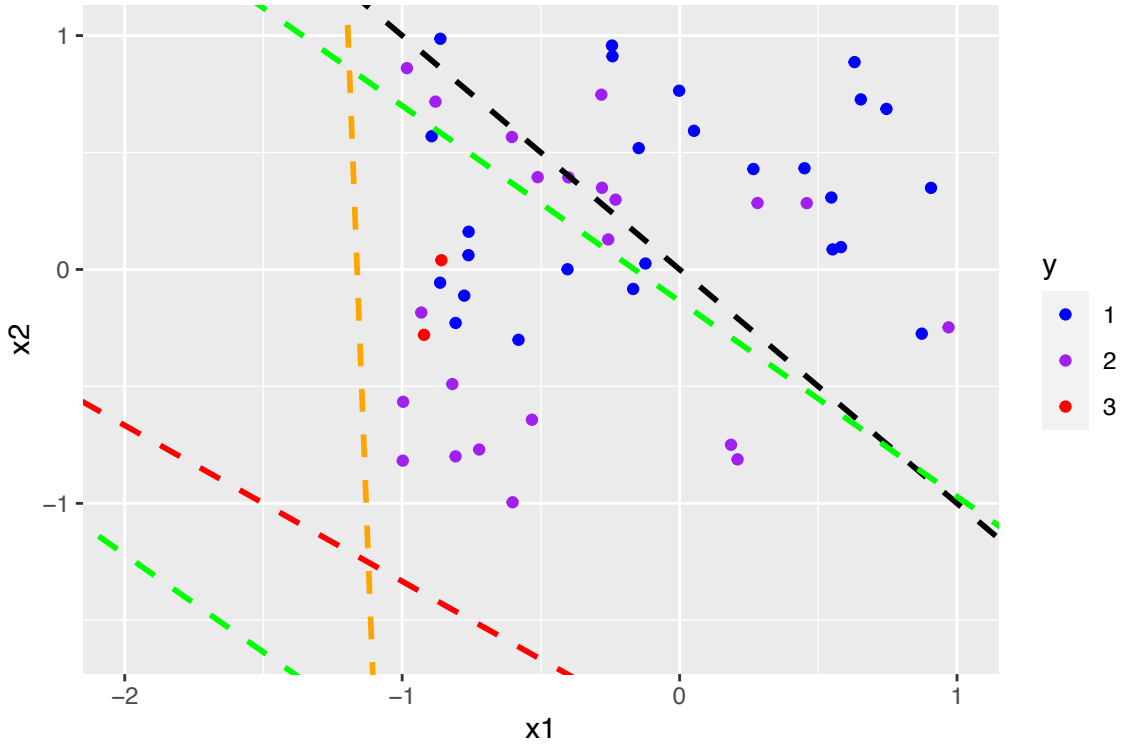


The Proportional Odds Model Leverages Data From More Common Outcomes

Green dashed lines: estimated decision boundaries from proportional odds model. (Orange decision boundary is from logistic regression on rare class, same as before.)

Proportional odds model does seem to yield a much better estimate of the rare decision boundary!

Also yields better estimates of $\mathbb{P}(y = 3 \mid \boldsymbol{x})$.



Problem solved? Not quite…

What's Wrong With Just Using The Proportional Odds Model? It's Unrealistically Rigid

Proportional odds model assumes that the β vector associated with each decision boundary is identical. Previous slide: when an outcome is rare, this is an improvement over separate logistic regressions to estimate each decision boundary.

But it's still unrealistically inflexible. **What if individual features have different influences on the decision boundaries at different levels?**

> In **online marketing**, it could be that some ad features make the ad "flashier" and increase the probability of a click, but are not predictive of the probability of purchasing.
>
> Students may place different weights on factors when **deciding whether to pursue graduate school vs. undergrad** (may have more appealing alternatives to additional schooling, may have different financial constraints, etc.)

What's Wrong With Just Using The Proportional Odds Model? It's Unrealistically Rigid

Proportional odds model assumes that the β vector associated with each decision boundary is identical. Previous slide: when an outcome is rare, this is an improvement over separate logistic regressions to estimate each decision boundary.

But it's still unrealistically inflexible. **What if individual features have different influences on the decision boundaries at different levels?**

> In **online marketing**, it could be that some ad features make the ad "flashier" and increase the probability of a click, but are not predictive of the probability of purchasing.
>
> Students may place different weights on factors when **deciding whether to pursue graduate school vs. undergrad** (may have more appealing alternatives to additional schooling, may have different financial constraints, etc.)

Ideally, we'd like the more common decision boundaries to *inform* the estimation of the rare decision boundaries without imposing exact equality.

Outline

1 Background and Motivation

2 PRESTO!

3 Simulation Studies

PRESTO!

Proportional odds model is typically estimated by maximum likelihood:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\alpha}\in\mathbb{R}^{K-1}}{\arg\min} \left\{ -\sum_{i=1}^{n} \log\left[F\left(\alpha_{y_i} + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right) - F\left(\alpha_{y_i-1} + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right) \right] \right\},$$

where $F(t) = \exp\{t\}/(1+\exp\{t\})$ is the standard logistic cdf.

(Interpretation: $F\left(\alpha_k + \boldsymbol{\beta}^\top \boldsymbol{x}\right) = \mathbb{P}(y \le k \mid \boldsymbol{x})$, so
$F\left(\alpha_k + \boldsymbol{\beta}^\top \boldsymbol{x}\right) - F\left(\alpha_{k-1} + \boldsymbol{\beta}^\top \boldsymbol{x}\right) = \mathbb{P}(y = k \mid \boldsymbol{x})$.)

PRESTO!

Proportional odds model is typically estimated by maximum likelihood:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\alpha}\in\mathbb{R}^{K-1}}{\arg\min} \left\{ -\sum_{i=1}^{n} \log\left[F\left(\alpha_{y_i} + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right) - F\left(\alpha_{y_i-1} + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right)\right]\right\},$$

where $F(t) = \exp\{t\}/(1+\exp\{t\})$ is the standard logistic cdf.

(Interpretation: $F\left(\alpha_k + \boldsymbol{\beta}^\top \boldsymbol{x}\right) = \mathbb{P}(y \le k \mid \boldsymbol{x})$, so $F\left(\alpha_k + \boldsymbol{\beta}^\top \boldsymbol{x}\right) - F\left(\alpha_{k-1} + \boldsymbol{\beta}^\top \boldsymbol{x}\right) = \mathbb{P}(y = k \mid \boldsymbol{x})$.)

PRESTO allows $\boldsymbol{\beta}$ to differ at each decision boundary, but imposes an ℓ_1 penalty on the differences between coefficients corresponding to the same feature in adjacent decision boundaries:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^{p\times(K-1)},\boldsymbol{\alpha}\in\mathbb{R}^{k-1}}{\arg\min} \left\{ -\sum_{i=1}^{n} \log\left[F\left(\alpha_{y_i} + \boldsymbol{\beta}_{y_i}^\top \boldsymbol{x}_i\right) - F\left(\alpha_{y_i-1} + \boldsymbol{\beta}_{y_i-1}^\top \boldsymbol{x}_i\right)\right] + \lambda \sum_{j=1}^{p}\sum_{k=2}^{K-1} \left|\beta_{jk} - \beta_{j,k-1}\right| \right\}$$

PRESTO!

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^{p\times(K-1)},\boldsymbol{\alpha}\in\mathbb{R}^{k-1}}{\arg\min}\left\{-\sum_{i=1}^{n}\log\left[F\left(\alpha_{y_i}+\boldsymbol{\beta}_{y_i}^{\top}\boldsymbol{x}_i\right)-F\left(\alpha_{y_i-1}+\boldsymbol{\beta}_{y_i-1}^{\top}\boldsymbol{x}_i\right)\right]+\lambda\sum_{j=1}^{p}\sum_{k=2}^{K-1}\left|\beta_{jk}-\beta_{j,k-1}\right|\right\}$$

Inspired by the fused lasso (Tibshirani et. al. 2005): places an ℓ_1 penalty not on the coefficients themselves, but on differences between adjacent coefficients.

ℓ_1 penalty encourages coefficients for adjacent decision boundaries to be similar $\Longrightarrow$ decision boundaries with abundant data influence the estimated rare decision boundaries. (Improves estimation of rare decision boundaries compared to logistic regression.)

But PRESTO has flexibility to allow differences in decision boundaries if observed data suggests it would help (improvement over proportional odds).

Makes sense if we assume that differences between adjacent coefficient vectors are (approximately) sparse.

Outline

Simulation Studies: Setup

500 simulations using $n = 2500$, $p = 10$, and $K = 4$ ordered responses:

Draw $\boldsymbol{X} \in [-1,1]^{n \times p}$, where $X_{ij} \sim \text{Uniform}(-1,1)$ for all i, j.

Generate $\boldsymbol{y} \in \mathbb{R}^n$ using a relaxation of proportional odds:

$$\log\left(\frac{\mathbb{P}\left(y \leq k \mid \boldsymbol{x}\right)}{\mathbb{P}\left(y > k \mid \boldsymbol{x}\right)}\right) = \alpha_k + \boldsymbol{\beta}_k^\top \boldsymbol{x}, \qquad k \in \{1,\ldots,K-1\},$$

where $\boldsymbol{\alpha} = (0,4,6)$ (first two classes are common, last class is rare), $\boldsymbol{\beta}_1 = (1,\ldots,1)^\top$, and $\boldsymbol{\beta}_k = \boldsymbol{\beta}_{k-1} + \boldsymbol{\psi}_k$ for random vectors $\boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_{K-1} \in \mathbb{R}^p$ (probability distributions specified later).

Estimate rare class probabilities by logistic regression on rare class, proportional odds, and PRESTO (penalty λ selected by cross-validation).

Calculate MSE of estimated rare class probabilities.

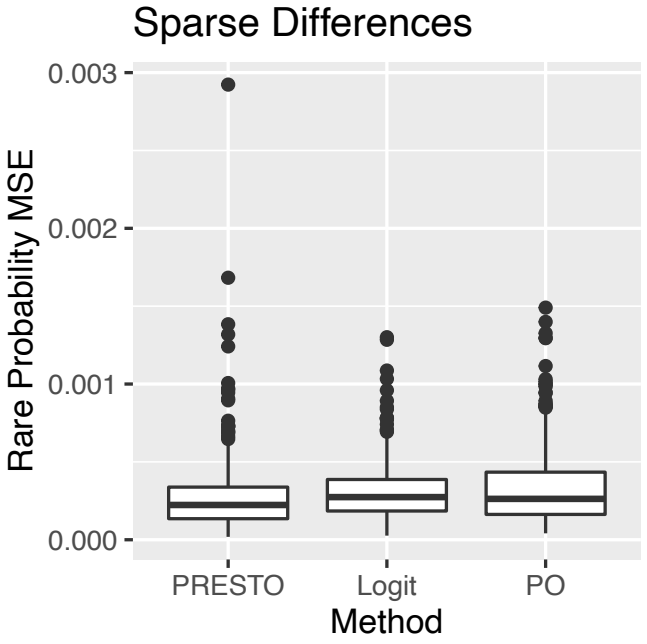
Simulation 1: Sparse Differences

$\boldsymbol{\beta}_k = \boldsymbol{\beta}_{k-1} + \boldsymbol{\psi}_k$, where
$\boldsymbol{\beta}_1 = (1,\ldots,1)^\top$ and

$$\psi_{jk} = \begin{cases} 0, & \text{with probability } 2/3, \\ 0.5, & \text{with probability } 1/6, \\ -0.5, & \text{with probability } 1/6, \end{cases} \qquad j \in \{1,\ldots,p\}.$$

(Should be a favorable setting for PRESTO due to sparsity.)

The results suggest that PRESTO does in fact estimate the rare probabilities more accurately!



Simulation 2: Dense (Approximately Sparse) Differences

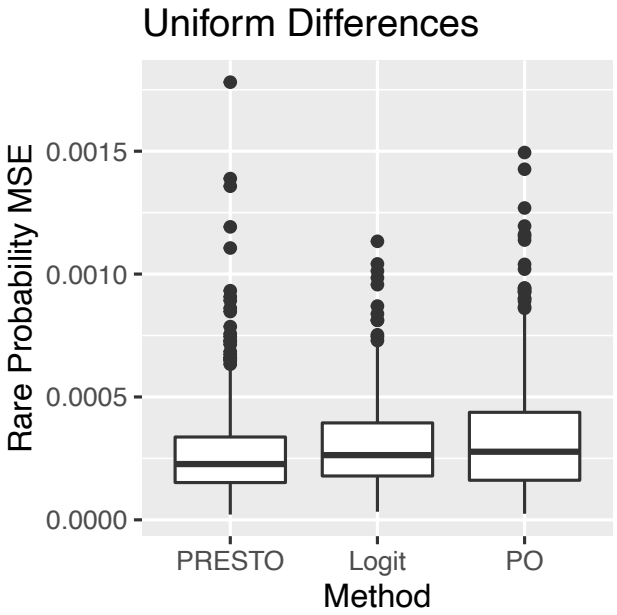
$\boldsymbol{\beta}_k = \boldsymbol{\beta}_{k-1} + \boldsymbol{\psi}_k$, where
$\boldsymbol{\beta}_1 = (1,\ldots,1)^\top$ and

$$\psi_{jk} \sim \text{Uniform}(-0.5, 0.5), \quad j \in \{1,\ldots,p\}.$$

$\boldsymbol{\psi}_k$ are not sparse $\Longrightarrow$ should be harder for PRESTO.

But $\boldsymbol{\psi}_k$ can be considered "approximately" sparse (some small entries approximately equal to 0, limited number of large entries that are important to account for).

PRESTO still outperforms logistic regression and proportional odds!



Summary

Binary classifiers struggle to estimate rare probabilities (class imbalance).

If there are ordinal outcomes, a decision boundary with abundant data nearby can be leveraged to improve estimation of rare decision boundary (and rare probabilities),

Proportional odds model allows this, but imposes exactly equality of $\boldsymbol{\beta}$ vectors (unrealistically rigid).

PRESTO relaxes proportional odds, allowing $\boldsymbol{\beta}$ vectors to differ but imposes ℓ_1 penalty on differences.

This allows for best of both worlds: learn from abundant decision boundaries, but flexibly adapt for different decision boundaries between different outcomes.

References

G. James, D. Witten, T. Hastie, & R. Tibshirani. *An introduction to statistical learning with applications in R* (Vol. 112, p. 18). New York: Springer, 2021.

P. McCullagh. Regression models for ordinal data. J*ournal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.