Calibrating Classification Probabilities with Restricted Polynomial Regression

Yongqiao Wang[1] Lishuai Li[2] Chuangyin Dang[2]

[1]School of Finance
Zhejiang Gongshang University

[2]Department of Systems Engineering and Engineering Management
City University of Hong Kong

UNF, Oct 18, 2019

Yongqiao Wang, Lishuai Li, Chuangyin Dang: Calibrating Classification Probabilities with Shape-Restricted Polynomial Regression. IEEE Trans. Pattern Anal. Mach. Intell. 41(8): 1813-1827 (2019)

Outline

Estimating membership probability in classification

Problem

- features $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ and label $Y \in \{0, 1\}$
- objective: an estimate of $g(\mathbf{x}) := \mathbb{P}\{Y = 1 | \mathbf{X} = \mathbf{x}\}$
- data: $\mathcal{D}_N = \{(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_N, Y_N)\}$

Estimating membership probability in classification

Problem

- features $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ and label $Y \in \{0, 1\}$
- objective: an estimate of $g(\mathbf{x}) := \mathbb{P}\{Y = 1 | \mathbf{X} = \mathbf{x}\}$
- data: $\mathcal{D}_N = \{(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_N, Y_N)\}$

Applications

- finance: determine the offered rate for a credit applicant
- handwritten character recognition: from the probability of each symbol to the probability of several symbols
- medicine: decide which therapy to give a patient
- ...

Bayesian on $\mathbf{x}$

One straightforward strategy is by the Bayesian rule

$$\begin{aligned}
&\mathbb{P}\{Y=1|\mathbf{X}=\mathbf{x}\}\\
=&\frac{\mathbb{P}\{\mathbf{X}=\mathbf{x}|Y=1\}\times\mathbb{P}\{Y=1\}}{\mathbb{P}\{\mathbf{X}=\mathbf{x}|Y=1\}\times\mathbb{P}\{Y=1\}+\mathbb{P}\{\mathbf{X}=\mathbf{x}|Y=0\}\times\mathbb{P}\{Y=0\}}
\end{aligned}$$

Bayesian on $\mathbf{x}$

One straightforward strategy is by the Bayesian rule

$$\mathbb{P}\{Y=1|\mathbf{X}=\mathbf{x}\} = \frac{\mathbb{P}\{\mathbf{X}=\mathbf{x}|Y=1\}\times\mathbb{P}\{Y=1\}}{\mathbb{P}\{\mathbf{X}=\mathbf{x}|Y=1\}\times\mathbb{P}\{Y=1\}+\mathbb{P}\{\mathbf{X}=\mathbf{x}|Y=0\}\times\mathbb{P}\{Y=0\}}$$

Difficult

It relies on the distributions of

$$\mathbf{X}|Y=1, \quad \mathbf{X}|Y=0, \quad Y$$

It is a big challenge for modeling $\mathbf{X}|Y=1$ and $\mathbf{X}|Y=0$, because commonly $\mathbf{X}$ is a mix of discrete and continuous random variables.

Classification

In a typical classification model, the label prediction for $\mathbf{x}$ is

$$l(s(\mathbf{x})) = \begin{cases} 1 & s(\mathbf{x}) \geq \mathfrak{h} \\ 0 & s(\mathbf{x}) < \mathfrak{h} \end{cases}$$

Two parts

- a scoring function $s : \mathcal{X} \rightarrow \mathbb{R}$
- a threshold function $l : \mathbb{R} \rightarrow \{0, 1\}$

Estimating membership probability with two steps

The prediction for the class-1 probability conditional on feature $\mathbf{x}$ is

$$\mathbb{P}\{Y = 1|\mathbf{X} = \mathbf{x}\} = f(s(x)) \tag{1}$$

Two parts

- a scoring function $s : \mathcal{X} \to \mathbb{R}$. Well studied, e.g. NN and SVM.
- a calibrating function $f : \mathbb{R} \to [0, 1]$. The focus of this paper!

Probability calibration

- score $S \in \mathcal{S} := s(\mathcal{X})$ and label $Y \in \{0, 1\}$
- objective: an estimate of $f(s) = \mathbb{P}\{Y = 1 | S = s\}$
- samples $D_N = \{(S_1, Y_1), \cdots (S_N, Y_N)\}$ where $S_n = s(X_n)$

Advantages

This strategy has two obvious advantages

- utilize nonlinear discriminant power of state-of-the-art classification models
- simplify from d dimensions ($\mathbf{x}$) to one dimension (s)

Advantages

This strategy has two obvious advantages

- utilize nonlinear discriminant power of state-of-the-art classification models
- simplify from d dimensions ($\mathbf{x}$) to one dimension (s)

Why this strategy can work?

A good scoring function is expected to satisfy:

$$\mathbb{P}\{Y = 1|\mathbf{X} = \mathbf{x}_1\} \geq \mathbb{P}\{Y = 1|\mathbf{X} = \mathbf{x}_2\}, \forall s(\mathbf{x}_1) > s(\mathbf{x}_2) \tag{2}$$

Outline

Calibration models

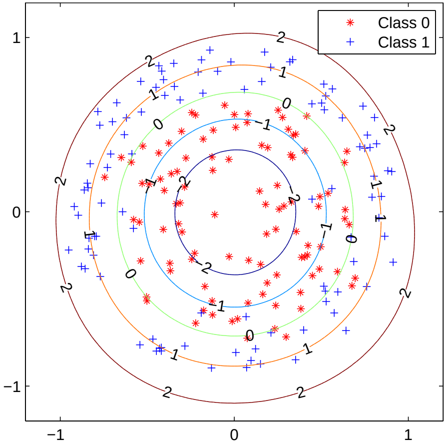
Name	Model	Function
Individual		
Platt	Logit regression	Sigmoid
HistBin	Histogram binning	Piecewise constant
IsoReg	Isotonic regression	Stepwise constant
NearIso	Nearly isotonic regression	Piecewise constant
LiTE	ℓ_1-linear trend filtering	Piecewise linear
ACP	Adaptive calibration	Piecewise constant
SmoIsoReg	Isotonic splines interpolation	Cubic splines
RPR(this paper)	Restricted polynomial regression	Polynomial
Ensemble		
BBQ	Ensemble of HistBin	Piecewise constant
ENIR	Ensemble of NearIso	Piecewise constant
ELITE	Ensemble of LiTE	Piecewise linear

Experiment - Toy Data

The toy data were generted

$$\mathbf{X} = (R\sin\theta, R\cos\theta)'$$
$$\theta \sim \text{Unif}(0, 2\pi)$$
$$R|Y = 0 \sim \text{Beta}(2, 5)$$
$$R|Y = 1 \sim \text{Beta}(5, 2)$$

Classifier: SVM with RBF kernel ($\sigma^2 = 1$)
Training size: 100 for each class
Test size: 200,000 for each class



How to measure calibrating performance?

- Calibrating performance can be measured by

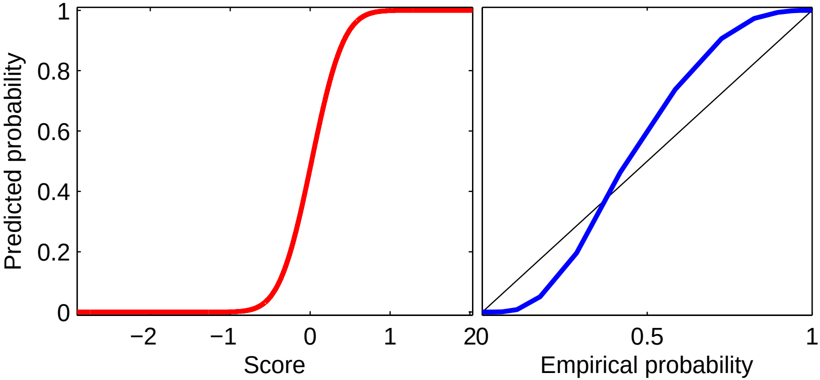
$$\|f - \hat{f}\|$$

$f = \mathbb{P}\{Y = 1|S = s\}$: the true calibrating function (unobservable)
$\hat{f} = \hat{\mathbb{P}}\{Y = 1|S = s\}$: the predicted calibrating function

- Even when the distribution of $(\mathbf{X}, Y)$ is known, commonly it is hard to derive f because of the complexity of s
- In this paper, the empirical ('true') function is obtaine by two steps
  1. All test scores are sorted in ascending order, and are partitioned into B subsets of equal frequency, called bins.
  2. For a test sample $S = s$, the prediction for $\mathbb{P}\{Y = 1|S = s(\mathbf{x})\}$ is the fraction of positive samples in the bin that includes s.

M1 Platt - Platt 1999

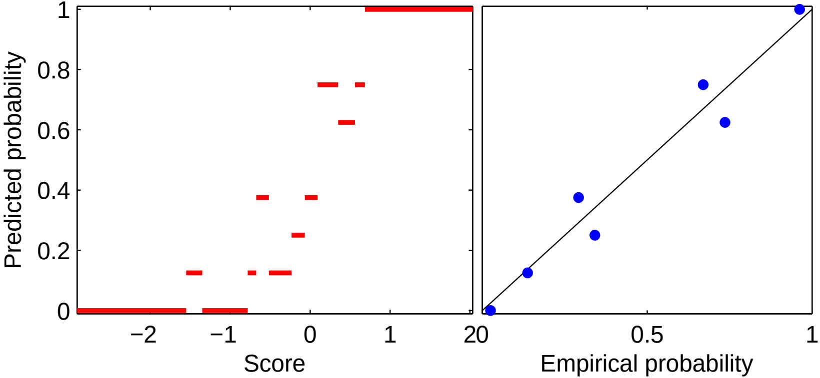
This model fits the training data with logit regression.



Disadvantage: the assumption of the sigmoid functional form.

M2 HistBin - Zadrozny and Elkan 2001

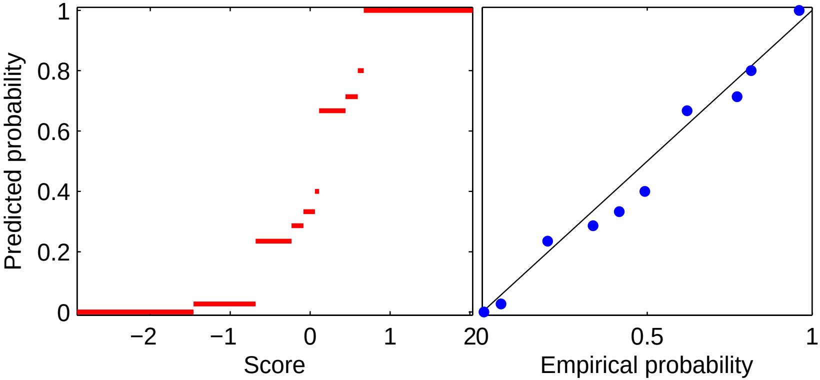
HistBin with $B = 25$



Disadvantages: (1) not increasing; (2) not continuous

M3 IsoReg - Zadrozny and Elkan 2002

This model fits the training data with isotonic regression.

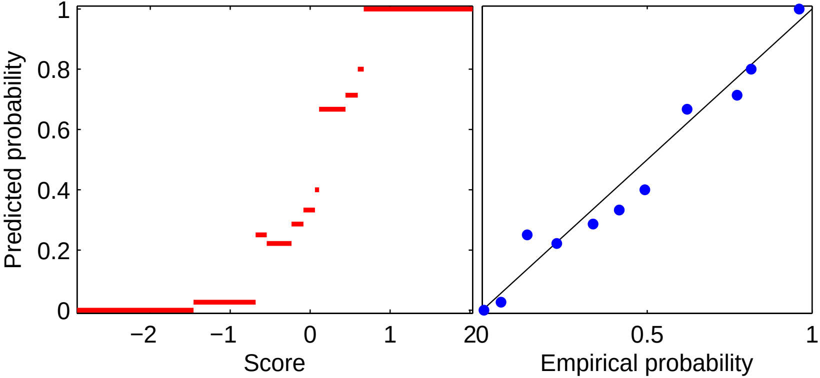


Disadvantage: not continuous

M4 NearISO - Naeini and Cooper 2018

This model fits the training data with nearly isotonic regression.

NearISO with $\lambda = 2$

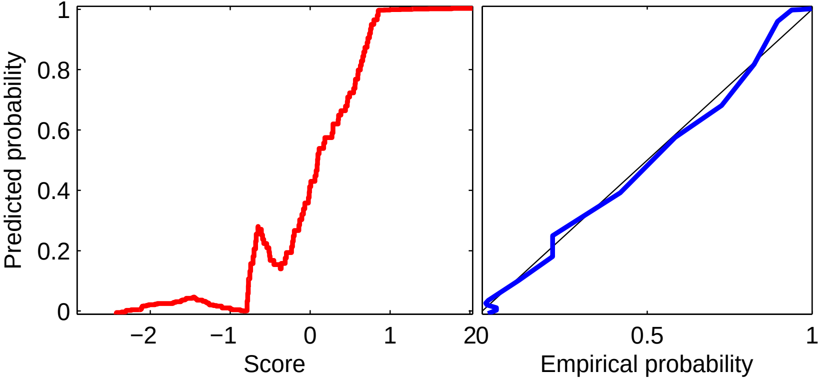


Disadvantages: (1) not increasing; (2) not continuous

M5 LiTE - Naeini and Cooper 2018

This model fits the training samples with ℓ_1 (linear) trend filtering signal approximation.

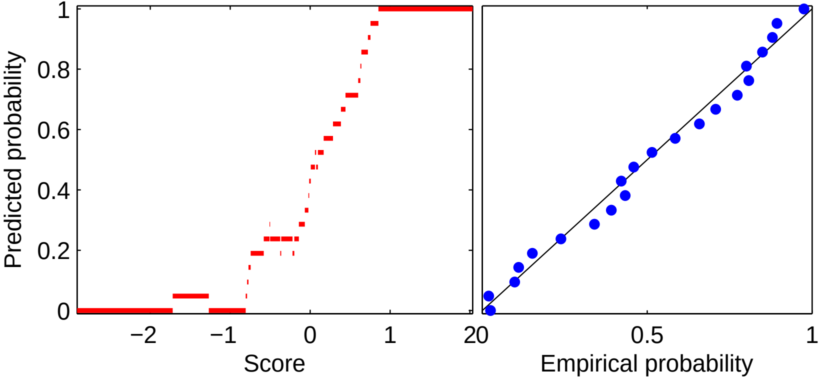
LiTE with $\lambda = 2$



Disadvantage: not increasing

M6 ACP - Zadrozny and Elkan 2001

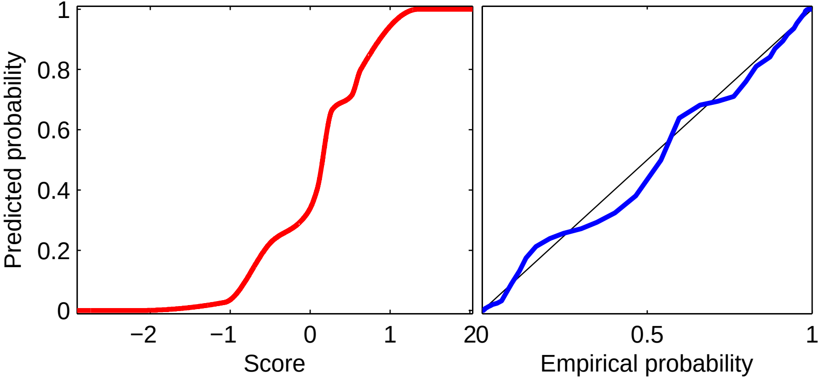
In the model the predicted probability for **x** is the percent of class 1 among its neighboring samples.



Disadvantages: (1) not increasing; (2) not continuous

M7 SMOIsoReg - Jiang et al. 2011

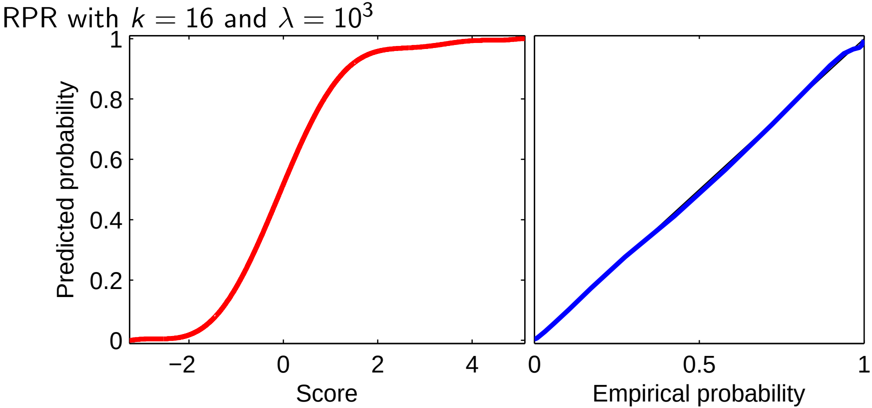
This model fits the training samples with increasing cubic splines regression.



Four requirements

1. increasing
2. continuous
3. universally flexible
4. computationally tractable

Estimated calibrating function from the proposed model



Qualitative comparison

Name	Flexibility	Monotonicity	Continuousness	Complexity
Individual				
Platt	-	+	+	$O(NT)$
HistBin	+	-	-	$O(N \log N)$
IsoReg	+	+	-	$O(N \log N)$
NearIso	o	-	-	$O(N \log N)$
LiTE	o	-	+	$O(N \log N)$
ACP	o	-	-	$O(N \log N)$
SmoIsoReg	-	+	+	$O(N^2)$
RPR(this paper)	+	+	+	$O(N^2)$
Ensemble				
BBQ	+	-	-	$O(N \log N)$
ENIR	o	-	-	$O(N^2)$
ELITE	o	-	+	$O(N \log N)$

+:satisfied, -: unsatisfied, o:unknown.

Outline

Framework

This model estimates the calibrating function f in the following framework

$$\min_{f \in \mathcal{F}} \quad \frac{1}{N} \sum_{n=1}^{N} [f(s_n) - y_n]^2 \tag{3}$$

where $\mathcal{F}$ is the family of continuous calibrating functions

$$\mathcal{F} := \left\{ f \in \mathcal{C}[\underline{s}, \overline{s}] \;\middle|\; \begin{array}{l} f(\underline{s}) \geq 0, \quad f(\overline{s}) \leq 1 \\ f(s) \text{ is non-decreasing over } [\underline{s}, \overline{s}] \end{array} \right\}. \tag{4}$$

Polynomial fitting

This model approximates f with a degree-k polynomial

$$f(s) = a_0 + a_1 s + \cdots + a_k s^k = \sum_{\ell=0}^{k} a_\ell s^\ell. \tag{5}$$

So

$$\begin{aligned} & f(s) = \sum_{\ell=0}^{k} a_\ell s^\ell \text{ is non-decreasing over } [\underline{s}, \bar{s}] \\ \Leftrightarrow & f'(s) = \sum_{\ell=1}^{k} a_\ell \ell s^{\ell-1} \geq 0, \forall s \in [\underline{s}, \bar{s}] \end{aligned}$$

Model

Thus the calibration problem becomes a semi-infinite program

$$\min_{\mathbf{a}\in\mathbb{R}^{k+1}} \quad \frac{1}{N}\sum_{n=1}^{N}\left[\sum_{\ell=0}^{k} a_\ell s_n^\ell - y_n\right]^2 \tag{6a}$$

$$s.t. \quad \sum_{\ell=0}^{k} a_\ell \underline{s}^\ell \geq 0, \quad \sum_{\ell=0}^{k} a_\ell \overline{s}^\ell \leq 1 \tag{6b}$$

$$\sum_{\ell=1}^{k} a_\ell \ell s^{\ell-1} \geq 0, \quad \forall s \in [\underline{s}, \overline{s}] \tag{6c}$$

$$\sum_{\ell=0}^{k} |a_\ell| \leq \lambda. \tag{6d}$$

This requirement includes uncountably infinite number of inequality constraints.

Lemma 1 - Nesterov 2000

Consider the polynomial $p(s) = a_0 + a_1 s + \cdots + a_k s^k$, $s \in [\underline{s}, \overline{s}]$.
(1) When k is even, e.g. $k = 2k_1$, $k_1 \in \mathbb{N}$, $p(s)$ is nonnegative on the closed interval $[\underline{s}, \overline{s}]$, if and only if there exist positive semidefinite real symmetric matrices $\mathbf{U} \in \mathbb{R}^{(k_1+1)\times(k_1+1)}$ and $\mathbf{V} \in \mathbb{R}^{k_1 \times k_1}$ satisfying

$$\begin{aligned} a_\ell = & \langle \mathbf{H}_{k_1+1,\ell+2}, \mathbf{U} \rangle + \langle -\underline{s}\overline{s}\mathbf{H}_{k_1,\ell+2}\mathbb{I}_{\{\ell \leq 2k_1-2\}} \\ & + (\underline{s} + \overline{s})\mathbf{H}_{k_1,\ell+1}\mathbb{I}_{\{1 \leq \ell \leq 2k_1-1\}} - \mathbf{H}_{k_1,\ell}\mathbb{I}_{\{\ell \geq 2\}}, \mathbf{V} \rangle \end{aligned} \tag{7}$$

for all $\ell = 0, \cdots, 2k_1$.
(2) When k is odd, $k = 2k_1 - 1$, $k_1 \in \mathbb{N}$, $p(t)$ is nonnegative on $[\underline{s}, \overline{s}]$, if and only if there exist positive semidefinite real symmetric matrices $\mathbf{U} \in \mathbb{R}^{k_1 \times k_1}$ and $\mathbf{V} \in \mathbb{R}^{k_1 \times k_1}$ satisfying

$$\begin{aligned} a_\ell = & \langle -\underline{s}\mathbf{H}_{k_1,\ell+2}\mathbb{I}_{\{\ell \leq 2k_1-2\}} + \mathbf{H}_{k_1,\ell+1}\mathbb{I}_{\{\ell \geq 1\}}, \mathbf{U} \rangle \\ & + \langle \overline{s}\mathbf{H}_{k_1,\ell+2}\mathbb{I}_{\{\ell \leq 2k_1-2\}} - \mathbf{H}_{k_1,\ell+1}\mathbb{I}_{\{\ell \geq 1\}}, \mathbf{V} \rangle \end{aligned} \tag{8}$$

for all $\ell = 0, \cdots, 2k_1 - 1$.

Hankel matrix

Let $\mathbf{H}_{n,\ell}$ be the $n \times n$ Hankel matrix with row-i column-j element

$$H_{n,\ell}^{ij} := \begin{cases} 1, & i + j = \ell \\ 0, & \text{otherwise}. \end{cases} \tag{9}$$

In case of even k

Let $k_1 = k/2$. The coefficients **a** can be obtained by solving the following semidefinite program

$$\min_{\mathbf{a},\mathbf{U},\mathbf{V}} \sum_{n=1}^{N} \left[\sum_{\ell=0}^{k} a_\ell s_n^\ell - y_n \right]^2 \tag{10a}$$

$$s.t. \sum_{\ell=0}^{k} a_\ell \underline{s}^\ell \geq 0, \quad \sum_{\ell=0}^{k} a_\ell \overline{s}^\ell \leq 1 \tag{10b}$$

$$\begin{aligned} \ell a_\ell = & \langle -\underline{s}\mathbf{H}_{k_1,\ell+1}\mathbb{I}_{\{\ell\leq 2k_1-1\}} + \mathbf{H}_{k_1,\ell}\mathbb{I}_{\{\ell\geq 2\}}, \mathbf{U}\rangle \\ & + \langle \overline{s}\mathbf{H}_{k_1,\ell+1}\mathbb{I}_{\{\ell\leq 2k_1-1\}} - \mathbf{H}_{k_1,\ell}\mathbb{I}_{\{\ell\geq 2\}}, \mathbf{V}\rangle \\ & \forall \ell = 1, \cdots, 2k_1 \end{aligned} \tag{10c}$$

$$\sum_{\ell=0}^{k} |a_\ell| \leq \lambda \tag{10d}$$

$$\mathbf{a} \in \mathbb{R}^{2k_1+1}, \mathbf{U}, \mathbf{V} \in \mathbb{R}_+^{k_1 \times k_1}. \tag{10e}$$

In case of odd k

Let $k_1 = (k+1)/2$. The coefficients $\mathbf{a}$ can be obtained by solving the following semidefinite program

$$\min_{\mathbf{a},\mathbf{U},\mathbf{V}} \sum_{n=1}^{N} \left[\sum_{\ell=0}^{k} a_\ell s_n^\ell - y_n \right]^2 \tag{11a}$$

$$s.t. \sum_{\ell=0}^{k} a_\ell \underline{s}^\ell \geq 0, \quad \sum_{\ell=0}^{k} a_\ell \overline{s}^\ell \leq 1 \tag{11b}$$

$$\begin{aligned} \ell a_\ell = {} & \langle \mathbf{H}_{k_1+1,\ell+1}, \mathbf{U} \rangle + \langle -\underline{s}\overline{s}\mathbf{H}_{k_1,\ell+1}\mathbb{I}_{\{\ell \leq 2k_1-3\}} \\ & + (\underline{s}+\overline{s})\mathbf{H}_{k_1,\ell}\mathbb{I}_{\{1\leq \ell \leq 2k_1-2\}} - \mathbf{H}_{k_1,\ell-1}\mathbb{I}_{\{\ell\geq 2\}}, \mathbf{V} \rangle \\ & \forall \ell = 1, \cdots, 2k_1 - 1 \end{aligned} \tag{11c}$$

$$\sum_{\ell=0}^{k} |a_\ell| \leq \lambda \tag{11d}$$

$$\mathbf{a} \in \mathbb{R}^{2k_1}, \mathbf{U} \in \mathbb{R}_+^{k_1 \times k_1}, \mathbf{V} \in \mathbb{R}_+^{(k_1-1)\times(k_1-1)}. \tag{11e}$$

Computational burden

- Semidefinite program is a generic convex optimization that be efficiently solved by off-the-shelf toolboxes, e.g. CVX.
- Regularly the computational cost of this estimation is as cheap as $O(N^2)$.

High-degree polynomials notoriously overfit training samples

- Nearly-perfect fit for training samples
- Wide fluctuation between training samples

High-degree polynomials notoriously overfit training samples

- Nearly-perfect fit for training samples
- Wide fluctuation between training samples

How does this model successfully suppress overfitting?

- a sufficient implement of the requirement of increasingness
- the smoothness from the regularization $\sum_{\ell=0}^{k} |a_\ell| \leq \lambda$

Outline

1. Motivation
2. Related work
3. Methodology
4. Theoretical analysis
5. Experiments
   - Model comparison
   - Computational complexity
6. Furture directions

Universal flexibity

Let $\mathcal{P}_k$ be the set of restricted algebraic polynomials with degree $\leq k$

$$\mathcal{P}_k := \left\{ \begin{array}{l} P_k : [\underline{s}, \overline{s}] \to \mathbb{R} \\ P_k(s) = \sum_{\ell=0}^{k} a_\ell s^\ell \end{array} \middle| \begin{array}{l} \sum_{\ell=0}^{k} a_\ell \underline{s}^\ell \geq 0, \ \sum_{\ell=0}^{k} a_\ell \overline{s}^\ell \leq 1 \\ \sum_{\ell=1}^{k} \ell a_\ell s^{\ell-1} \geq 0, \forall s \in [\underline{s}, \overline{s}] \end{array} \right\}. \tag{12}$$

Universal flexibity

Let $\mathcal{P}_k$ be the set of restricted algebraic polynomials with degree $\leq k$

$$\mathcal{P}_k := \left\{ \begin{array}{l|l} P_k : [\underline{s}, \overline{s}] \to \mathbb{R} & \sum_{\ell=0}^{k} a_\ell \underline{s}^\ell \geq 0, \ \sum_{\ell=0}^{k} a_\ell \overline{s}^\ell \leq 1 \\ P_k(s) = \sum_{\ell=0}^{k} a_\ell s^\ell & \sum_{\ell=1}^{k} \ell a_\ell s^{\ell-1} \geq 0, \forall s \in [\underline{s}, \overline{s}] \end{array} \right\}. \tag{12}$$

Theorem (Universal flexibility)

$\cup_{k=1}^{\infty} \mathcal{P}_k$ is dense in $\mathcal{F}$ with respect to sup-norm, i.e. for any $f \in \mathcal{F}$,

$$\lim_{k \to \infty} \min_{P_k \in \mathcal{P}_k} \|f - P_k\|_\infty = 0. \tag{13}$$

Theorem (Universal statistical convergence)

If $\{k_N\}$ *and* $\{\lambda_N\}$ *satisfy*

$$\lambda_N \uparrow \infty,\ k_N \uparrow \infty,\ k_N/N \to 0 \tag{14}$$

then,
(a) $\{\hat{f}_N\}$ *is weakly universally consistent, i.e.*

$$\lim_{N\to\infty} \mathbb{E}\left\{ \int (f(s) - \hat{f}_N(s))^2 \mu(\mathrm{d}s) \right\} = 0 \tag{15}$$

for all distributions of (S, Y).
(b) $\{\hat{f}_N\}$ *is strongly universally consistent, i.e.*

$$\lim_{N\to\infty} \int (f(s) - \hat{f}_N(s))^2 \mu(\mathrm{d}s) = 0,\ \textit{with probability 1} \tag{16}$$

for all distributions of (S, Y).

Outline

Experiment - data

- two data from UCI
  - Adult: 14 features and 45,222 samples
  - Bank Marketing: 20 features and 45,211 samples
- two classifier
  - Logit regression
  - SVM with RBF kernel
- hyperparameters determination: 4-fold cross-validation
- training size: 200 or 500
- test size: all other samples
- repeatation rounds: 50

Performance measures

$$\text{ECE} = \sum_{i=1}^{K} |p_i - e_i|/K, \tag{17}$$

$$\text{MCE} = \max_{i=1,\cdots,K} |p_i - e_i| \tag{18}$$

p_i: predicted probability from training data.
e_i: predicted probability with HistBin from test data ('true probability').
$K = 100$ in this experiment.

Model comparison - scoring with logit regression

	Adult				Bank Marketing			
	$N = 200$		$N = 500$		$N = 200$		$N = 500$	
	MCE	ECE	MCE	ECE	MCE	ECE	MCE	ECE
Platt	8.918±2.337	3.734±0.989	7.953±3.158	4.179±1.185	9.107±3.143	3.175±1.588	8.576±3.545	4.130±1.271
Hist	11.038±2.981	5.931±1.690	7.041±2.498	2.504±0.834	13.922±4.083	4.858±0.832	8.354±2.575	3.882±0.921
Isoreg	7.500±2.318	4.074±1.309	7.091±2.625	<u>2.483</u>±0.643	9.064±2.076	3.139±1.195	7.130±2.800	3.005±0.470
NearIso	10.561±3.057	3.385±1.542	8.071±3.265	3.202±0.893	11.267±3.045	6.463±1.502	8.677±1.519	3.110±0.914
LiTE	6.978±3.273	3.073±1.032	5.453±2.480	2.789±0.979	7.902±2.775	2.629±1.465	6.287±2.295	2.899±0.844
ACP	9.875±2.917	3.639±1.304	8.025±3.165	3.164±1.524	9.277±3.486	4.034±1.146	8.173±1.732	4.705±1.644
SmoIsoReg	6.872±2.001	2.901±0.562	5.071±2.321	2.720±0.786	6.532±1.521	<u>2.612</u>±1.213	<u>4.569</u>±2.329	<u>2.044</u>±0.768
RPR	**4.291**±1.212	**1.677**±0.734	**3.615**±1.881	2.613±0.751	**4.817**±1.388	**2.539**±0.727	**4.351**±2.289	**1.952**±0.901
BBQ	<u>5.752</u>±2.981	3.643±0.735	5.468±3.264	2.557±1.266	10.820±2.405	2.938±1.179	6.641±3.510	2.329±1.069
ENIR	6.687±2.079	2.691±1.517	7.660±3.616	2.909±1.140	6.816±1.404	2.631±1.416	6.985±2.489	3.152±1.212
ELiTE	6.731±1.885	<u>2.492</u>±1.093	<u>4.110</u>±2.109	**2.244**±0.718	<u>6.300</u>±2.303	3.590±1.288	5.836±1.992	2.143±0.666

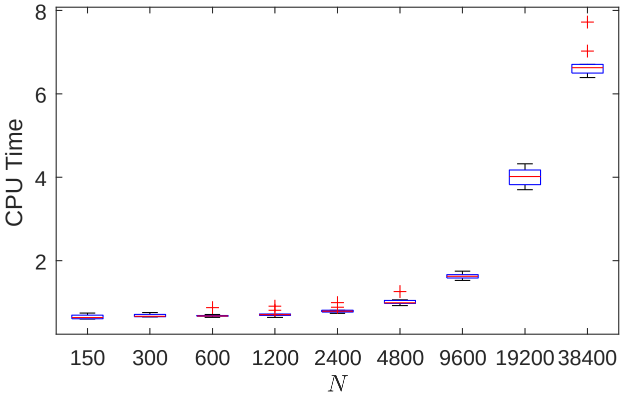
In each cell $a \pm b$: a is the average and b is the standard deviation. In each column, the best performance is in bold and the second best is underlined.

Model comparison - scoring with SVM

	Adult				Bank Marketing			
	$N = 200$		$N = 500$		$N = 200$		$N = 500$	
	MCE	ECE	MCE	ECE	MCE	ECE	MCE	ECE
Platt	8.684±2.538	4.971±1.956	8.442±3.866	4.043±3.795	7.061±2.990	5.500±1.438	6.536±4.277	4.594±2.825
HistBin	11.785±3.924	7.227±4.389	6.947±3.594	4.640±2.725	12.289±5.759	6.443±3.495	8.353±3.222	4.407±2.797
IsoReg	9.732±4.062	5.301±3.809	8.875±3.319	4.487±3.126	9.353±4.495	5.456±2.019	8.113±3.909	5.268±1.995
NearIso	13.381±4.558	8.299±2.772	10.543±3.948	6.259±3.864	11.901±4.470	4.275±2.357	5.262±3.821	3.962±2.548
LiTE	8.722±4.628	5.695±3.552	6.835±5.725	3.817±3.456	8.512±3.017	4.785±2.371	6.419±5.468	5.290±3.137
ACP	8.023±2.898	<u>3.017</u>±1.154	7.033±2.796	<u>2.628</u>±1.069	7.425±2.578	3.417±0.997	7.170±2.775	3.314±1.120
SmoIsoReg	<u>5.005</u>±2.385	3.543±1.129	<u>5.295</u>±2.122	2.820±1.127	<u>5.541</u>±2.718	**2.473**±1.663	5.968±2.715	3.071±1.236
RPR	**4.440**±1.601	**2.239**±0.932	**4.362**±1.965	**2.002**±1.219	**4.704**±2.761	<u>2.615</u>±1.038	<u>4.549</u>±1.828	<u>2.549</u>±1.250
BBQ	8.007±4.298	4.936±2.052	7.897±4.405	4.609±2.965	6.199±4.648	4.075±1.787	6.256±3.897	3.580±1.798
ENIR	9.314±4.981	4.303±1.768	8.115±3.399	4.373±2.841	7.574±2.351	4.041±2.037	6.154±2.738	3.036±1.642
ELITE	8.292±2.113	4.539±2.069	6.583±2.605	2.851±1.175	6.589±3.563	3.648±0.958	**4.138**±2.030	**2.354**±1.161

In each cell $a \pm b$: a is the average and b is the standard deviation. In each column, the best performance is in bold and the second best is underlined.

Computational time



Data: Adult. Classifier: SVM. PC: Core i5-2400 CPU @3.10GHz and 4GB RAM.

Outline

Future directions

1. one classifier $\Rightarrow$ multiple classifiers
2. binary classification $\Rightarrow$ multi-class classification

Thank you!