Highly Accurate Linear Classifier with Applications in Health Insurance Coverage

PhD Thesis Presentation

Songkomkrit Chaiyakan

Graduate School of Applied Statistics
National Institute of Development Administration

Introduction

Casual Explanation in Social Science Research

Logistic Regression and SEM

- Significance test of coefficients
- Mediators, moderators, confounders and covariates
- Independent variables measured on nominal scales require the excessive number of required dummy variables

Classification Algorithms in ML

- Decision tree: set of rules
- Neural network: nonlinear interaction as a result of a hidden layer

Performance Metrics

- Training/testing accuracy is not defined as their objective functions

Proposed Classification

- Developed from SVM through 0-1 MILP
- Count the number of misclassified instances through majority voting
- Ensure maximum accuracy without overfitting simply due to its linearity
- Serve no purpose of real-time analytics
- No consideration of interrelationship between contributing factors
- Trained on the entire survey data
  - All responses collected from different participants are of equal importance
  - No prediction about future health insurance coverage is made

Objectives

1. To propose a multiclass box classifier that yields highest training accuracy
2. To apply the proposed classification method to investigate significant factors, whether continuous or categorical, influencing health insurance coverage

Limitations

1. Nonlinear classification and logistic regression are beyond the scope of the study
   - No interaction between health insurance factors is investigated
   - Splitting values on any two factors should be independent
2. The health insurance sample data only includes Americans
3. Only three factors are preselected and investigated with a sample size of 100
   - Significantly long training time
   - Enormous space to store a branch-and-cut tree
   - Approximation algorithm is not developed in this dissertation
   - The model training lasts longer than a day
   - Nonetheless, the early-exit classifiers are more accurate and parsimonious than a Gini-based decision tree

Literature Review

Existing Analytical Techniques

Technique	**Reference**
Linear probability modeling	Cebula (2006)
Probit regression analysis	Mulenga et al. (2021)
Binomial logistic regression analysis	Markowitz et al. (1991) Dolinsky and Caputo (1997)
Multinomial logistic regression analysis	Jin et al. (2016)

Literature Review of Determinants

Technique	Population	Contributing factors
Linear (2006)	American (state-level)	+ income/F/$\geq$65; − wife/Hispanic
Probit (2021)	Zambian (psychological)	+ service/skilled/unskilled/rural (M); + marital union / clerical (F)
BLogit (1991)	American (18-24)	F > P (permanent); P < F (student)
BLogit (1997)	female American	− income/edu; + health (married); − employment (unmarried)
MLogit (2016)	Chinese (45 and over)	income/education/health/employment $\rightarrow$ rural+urban/private/pub+private

Decision Tree

Binary Tree

- Splitting value
  - Minimize the weighted average of the impurities of both child nodes by their number of training instances
- Impurity measure $\leftarrow$ Gini index
  - Gini index = Prob of a sample at a node being wrongly classified
- Categorical feature $\rightarrow$ One-hot encoding

Multiway Tree $\rightarrow$ Binary tree

Node having successors $V_1, V_2, V_3, \ldots$

- Left child = V_1
- Right child = $\neg V_1$ (negation)
  - Successors = $V_2, V_3, \ldots$

This recursive operation maintains the decision regions

SelectKBest (1/2)

Definition (Kullback-Leibler Distance)

The *Kullback-Leibler distance* $D(f||g)$ between two densities f and g is defined by

$$D(f||g) = \int f \log \frac{f}{g}.$$

Definition (Mutual Information)

The *mutual information* $I(X;Y)$ between two random variables with joint density $f(x,y)$ is defined as

$$I(X;Y) = D(f(x,y)||f(x)f(y)).$$

X and Y share no mutual information $\iff$ $I = 0$ $\iff$ X and Y are independent

SelectKBest (2/2)

Definition (Gamma function Γ and digamma function ψ)

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \quad \text{and} \quad \psi(z) = \frac{d}{dz} \log \Gamma(z)$$

- X = training variable and Y = discrete target or class
- X is continuous $\rightarrow$ Estimate $I(X;Y) \rightarrow k$-nearest neighbor method (Ross, 2014)
- The k-nearest neighbor of x^i of the same class has m_i instances of all classes
- There are N_i out of N that share the same class with x^i
- Compute $I_i = \psi(N) - \psi(N_i) + \psi(k) - \psi(m_i)$
- $I(X;Y) \approx$ Average of I_i over all training instances

Research Methods

Overview

1. Propose a multiclass box classifier which is able to predict continuous contributing factors, produces disconnected decision regions and provides minimum misclassification
2. Extend the classifier when certain features of training data are allowed to be categorical
3. Connect to a cloud virtual machine using secure shell (SSH) and install Python from source as well as CPLEX
4. Illustrate the use of the proposed classification method on the health insurance dataset
5. Compare multiple facets of results with the use of a decision tree
6. Back up the scripts and results to Oracle Cloud Infrastructure (OCI) Object Storage
7. Publish the project to GitHub

SSH Key Generation (1/2)

- SSH keys can be generated with the OpenSSH command `ssh-keygen` by using a native SSL/TLS library provided by an operating system
  - Secure Channel (Schannel) in Windows
  - OpenSSL in Linux
- In this dissertation, the SSH keys are created on a local computer with the elliptic-curve Ed25519 algorithm (Bernstein et al., 2012), proven to be faster and more efficient than the RSA algorithm (Rivest et al., 1978)

```
cd ~/.ssh
ssh-keygen -f <output_keyfile> -C <comment> -t ed25519
```

- A Google Cloud virtual machine requires the comment at the end of a public key file to be a Google username

SSH Key Generation (2/2)

- Different purposes → Different key pairs → Tightened security
  - The default private key `id_rsa` may not be used for authentication
- Specify a `Host` (named host) in the configuration file `~/.ssh/config`
  - `HostName`: IP address or domain name (can be omitted if = `Host`)
  - `User`: username
  - `IdentityFile`: path to private key file

```
Host <named_host>
    HostName <hostname>
    User <username>
    IdentityFile <private_keypath>
```

- Principle of least privilege (PoLP)
  - Loose private key permissions → Refused SSH connection (to prevent privilege escalation attacks)

SSH Key Generation on Linux

- Three POSIX permission levels: owner, group and other
- Each level is represented by three permission bits: read (r), write(w) and execute (x)
  - Usually rewritten in base 10, ranging from 0 to 7
- The `chmod` command is used to set all three levels of permission with three numerical digits

```
chmod 400 <private_key>
chmod 444 <public_key>
```

SSH Key Generation on Windows (1/4)

- Additional rights can be denied due to more fine-grained permission control
- General characteristics of generated SSH key pair
  - Hidden + No inherited NTFS permission + Nontransferable ownership
- Consider the following set $\mathcal{A}$ of access privileges
  - Content: read-only
  - Regular and extended attributes: read-only
  - Permissions: read-only
- Suppose a current user who creates and owns a key pair is the only administrator
  - This assumption is valid on, for example, a personal computer (PC)

Key Type	Key Owner	Administrators	SYSTEM	Everyone
Private	$\mathcal{A}$	$\mathcal{A}$	Deny All	–
Public	$\mathcal{A}$	$\mathcal{A}$	$\mathcal{A}$	$\mathcal{A}$

SSH Key Generation on Windows (2/4)

The command `icalcs` is used to manage file permissions

```
icacls <key> /inheritancelevel:d
icacls <key> /grant ${Env:USERNAME}:F Administrators:F SYSTEM:F
    Everyone:F
attrib +h <key>
icacls <key> /remove ${Env:USERNAME} Administrators SYSTEM Everyone
```

SSH Key Generation on Windows (3/4)

Permission	Description
WD	Write data or add file
AD	Append data or add subdirectory
WA	Write attributes
WEA	Write extended attributes
DE	Delete
WDAC	Write DAC (change permissions)
WO	Write owner (take ownership)

```
icacls <key> /deny "${Env:USERNAME}:(WD,AD,WA,WEA,DE,WDAC,WO)" "
    Administrators:(WD,AD,WA,WEA,DE,WDAC,WO)"
icacls <key> /grant ${Env:USERNAME}:R Administrators:R
```

SSH Key Generation on Windows (4/4)

Permission	Description
WD	Write data or add file
AD	Append data or add subdirectory
WA	Write attributes
WEA	Write extended attributes
DE	Delete
WDAC	Write DAC (change permissions)
WO	Write owner (take ownership)

```
icacls <private_key> /deny SYSTEM:F
icacls <public_key> /deny "SYSTEM:(WD,AD,WA,WEA,DE,WDAC,WO)" "
    Everyone:(WD,AD,WA,WEA,DE,WDAC,WO)"
icacls <public_key> /grant SYSTEM:R Everyone:R
```

VM Specifications

- Google Cloud compute engine
- 64-bit 8-vCPU 4-core CPU, 64 GB RAM and 250 GB SSD persistent disk
- Ubuntu Server 24.04 LTS
- Region `us-central1 (Iowa)` and zone `us-central1-f`
- Standard provisioning model (no VM preemption)
- Premium-tier network traffic (low latency)
- Reserved static external IPv4 address

SSH Key-Based Authentication on VM

- Password authentication should be disabled
- Modify `/etc/ssh/sshd_config`
  - Uncomment the line: `PasswordAuthentication no`
- Add a public key of a local computer to the key file `~/.ssh/authorized_keys`

```
echo <public_keyfile> >> ~/.ssh/authorized_keys
```

String Interning in Python

- Python interns strings (immutable objects) of the same value mainly through the function `_PyUnicode_InternInPlace()`
  - Only one copy is retain in memory
  - This function is defined in the source file `Objects/unicodeobject.c`
- Goals
  - To reduce memory usage
  - To speed up certain operations (e.g. equality comparison)
- The reference to all interned strings is stored in the per-interpreter dictionary `interned` initialized during the first invocation
- A debug build denies with an assertion (as opposed to a release build) the addition of a process-global interned string into the existing dictionary
  - No chance of getting a duplicate

Manual Python Installation (1/3)

Python 3.13.0 is built and installed (separate from the system Python 3.12.3 library)

- Manually built with GCC 15 experimental
- Debug build with shared library
- Global interpreter lock (GIL) still enabled to install `scikit-learn` successfully
- Both compilation and linking are optimized (PGO = Profile-guided optimization)
- Profiling is turned off (by default)
- The OpenSSL crypto policy `openssl.cnf` is respected by overriding the default Python cipher list
- Opt out of string interning to launch `JupyterLab` successfully

```
sed -i -e "s/assert(interned == NULL);/\/\/assert(interned == NULL)
    ;/g" Objects/unicodeobject.c
```

Manual Python Installation (2/3)

```
export PROFILE_TASK="-m test --pgo --timeout=300 -i test_embded"

cd <build_dir>

<source_dir>/configure --prefix=<install_dir> \
    --with-openssl=<openssl_rootdir> --with-openssl-rpath=auto \
    --enable-shared --enable-optimizations --with-lto=full \
    --with-pydebug --enable-loadable-sqlite-extensions \
    --with-computed-gotos --with-valgrind --with-dtrace \
    --with-system-libmpdec --with-system-expat \
    --with-ssl-default-suites=openssl

make -j<N>
make -j<N> install
```

Manual Python Installation (3/3)

```
export PATH="<install_dir>/bin:$PATH"
export LD_LIBRARY_PATH="<install_dir>/lib:${LD_LIBRARY_PATH}"
export LDFLAGS="-L<install_dir>/lib $LDFLAGS"

export PYTHONWARNINGS="ignore::DeprecationWarning"
```

Oracle Cloud Infrastructure (OCI)

Logical Concepts of Organization Management

- Tenancy: root container
  - During the signup process, a parent tenancy is provisioned and tied to a specified, unchangeable home region which is `ap-singapore-1` in this dissertation
  - Multiple child tenancies can be created and managed by the parent tenancy
- Compartment
  - Belong to a tenancy
  - Control access to cloud resources
  - Up to 6 levels are supported
  - It must be specified when a resource is created

OCI Object Storage

- Namespace: top-level container for all buckets and objects
  - It is unique to a tenant
  - It spans all compartments within a region
  - Although region-specific, its name remains the same across all regions
- Bucket: logical container unique in a namespace
- Object: any type of data along with its metadata stored in a bucket
- Tiers
  - Standard: high cost and no retention period
  - Archive: retention of at least 90 days with slow restoration
- OCI Object Storage supports auto-tiering, object versioning and multipart uploading
- Uncommitted of failed multipart uploads can be cleaned either manually or through a predefined lifecycle policy rule

Backup to OCI Object Storage

Free Tier

- A total of 20 GB in all tenancies is always free
- No upgrade to a paid account is required

Dissertation

- A bucket is created without auto-tiering and versioning
- Only a full backup of scripts and results is stored in OCI Object Storage
  - Backup size is very small
  - Avoid the possibility of a corrupted incremental or differential backup

OCI CLI Installation

```
~$ python3 -m venv <env_dir>
~$ source <env_dir>/bin/activate
(env_dir)$ pip3 install oci-cli
(env_dir)$ deactivate
```

- Executables (including `oci`) are in the `bin` directory
  - Add to `PATH`
- Libraries are in the `lib` directory
  - No need to add to `LD_LIBRARY_PATH`

OCI CLI Configuration

Interactive: `oci setup config`

Noninteractive

- Set the environment variable `OCI_CLI_RC_FILE` to the configuration file path

```
[DEFAULT]
user=<user>
fingerprint=<fingerprint>
key_file=<key_file>
tenancy=<tenancy>
region=ap-singapore-1
```

- A nondefault section should be specified via the `--profile` option in the CLI

OCI CLI Commands

```
oci os object sync -ns <namespace> -bn <bucket> --prefix <
    obj_prefix> --src-dir <src_dir> --delete

oci os object rename -ns <namespace> -bn <bucket> --name <obj_name>
     --new-name <obj_new_name>

oci os object delete -ns <namespace> -bn <bucket> --name <obj_name>

oci os object bulk-delete -ns <namespace> -bn <bucket> --prefix <
    obj_prefix>
```

Git Commands

Command	Description
`git clone`	Clean copy
`git pull`	Update with local changes kept
`git reset --hard`	Update with local changes discarded
`git clean -fdx`	Clean with untracked files and directories removed
`git push`	Remote update with local commits

Git Configuration

Interactive

```
git config --global user.name <username>
git config --global user.email <email_address>
```

Noninteractive

- Set the environment variable `GIT_CONFIG_GLOBAL` to the path to the Git global configuration file `.gitconfig`

```
[user]
name = <username>
email = <email_address>
```

GitHub Repository

Template GitHub Repository

- Available at `https://github.com/songkomkrit/phd-template`
- Very minimal with merely output files generated by a CPLEX optimizer

Up-to-date GitHub Repository

- Available at `https://github.com/songkomkrit/phd`
- Based on the template with additional outputs included

OPL Model Execution

```
oplrun -p <project_dir> 2>&1 | tee <log_file>
```

CPS ASEC Dataset

- CPS = Current Population Survey
- ASEC = Annual Social and Economic Supplement
- 2020 = Estimates of 2019
- Three levels: household / family / person (only used in the dissertation)
- This dissertation considers 184 independent variables (allocation flags excluded)
  - 66 continuous
  - 118 categorical (mostly nominal)

Categories of Health Factors

Topic	Subtopic	Num of Variables
Demographics	Individual characteristics	18
Basic CPS items	Edited labor force items	5
	Edited earnings items	4
	Labor force person recodes	16
Work experience	General	14
Income	Earnings	17
	Other income	80
	Non-cash benefits	6
	Supplemental poverty measure	6
	Tax model items	14
Poverty	Poverty	2
Health insurance	Health status	1
Supplemental poverty measure	SPM unit characteristics	1

Data Dictionary (1/3)

Age

- **Variable:** A_AGE
- **Topic:** Demographics
- **Subtopic:** Individual characteristics
- **Type:** Continuous
- **Range of values:** 00:85
- **Universe:** All persons

Values

00-79 = 0-79 years of age
80 = 80-84 years of age
85 = 85+ years of age

Data Dictionary (2/3)

Major labor force recode

- **Variable:** PEMLR
- **Topic:** Basic CPS items
- **Subtopic:** Labor force person recodes
- **Type:** Categorical (nominal)
- **Range of values:** 0:7
- **Universe:** All persons

Values

0 = NIU (not in universe)
1 = Employed - at work
2 = Employed - absent
3 = Unemployed - on layoff
4 = Unemployed - looking
5 = Not in labor force - retired
6 = Not in labor force - disabled
7 = Not in labor force - other

Data Dictionary (3/3)

Who received social security payments either for themselves or as combined payments with other family members?

- **Variable:** SS_YN
- **Topic:** Income
- **Subtopic:** Other income
- **Type:** Categorical (nominal)
- **Range of values:** 0:2
- **Universe:** All persons aged 15+

Values

0 = NIU (not in universe)
1 = Yes
2 = No

Types of Health Insurance Coverage

Private

- Employment-based plan: Employer-provided plan / Union-provided plan
- Direct-purchase plan

Government

- Medicare
- Medicaid / SCHIP (State Children's Health Insurance Plan)
- Military health care: TRICARE or CHAMPUS / CHAMPVA / VA
- State-specific plan
- IHS (Indian Health Service)

Uninsured

- IHS + no other type of health insurance

Scope of Study

- A group of infant born after the calendar year is excluded
- The combination of three following coverages is considered
  - Employment-based plan (GRP)
  - Direct-purchase plan (DIR)
  - Public health insurance (PUB)

Class	Code	GRP	DIR	PUB
0	NNN	No	No	No
1	NNY	No	No	Yes
2	NY_	No	Yes	Yes
		No	Yes	No
3	YNN	Yes	No	No
4	Y1Y	Yes	No	Yes
		Yes	Yes	Yes
		Yes	Yes	No

Exploratory Data Analysis (EDA)

- All infants born after calendar year are excluded in this study because they are not in the scope of health insurance coverage
- This results in 157,681 relevant survey participants
- The original dataset of size 237.4 MB in SAS7BDAT format can significantly be compressed into the feather and CSV formats of size 14.2 MB and 68.1 MB respectively
- A pandas accessor is used to compute the cross tabulation between a health factor (independent variable) and a combination of categorical insurance coverage types (dependent variable)

Cross Tabulation (1/3)

Variable	Insurance Coverage Type (GRP, DIR, PUB)				
	NNN	NNY	NY_	Y1Y	YNN
A_AGE: Age					
(-0.085, 8.5]	1,407	5,834	789	628	9,795
(8.5, 17.0]	1,557	6,237	1,079	770	11,822
(17.0, 25.5]	2,238	2,475	1,043	414	8,017
(25.5, 34.0]	2,635	2,749	1,082	594	10,611
(34.0, 42.5]	2,271	2,146	976	613	11,509
(42.5, 51.0]	2,109	2,171	1,157	518	12,081
(51.0, 59.5]	1,606	2,403	1,223	471	9,864
(59.5, 68.0]	1,028	4,854	2,313	2,090	6,097
(68.0, 76.5]	105	5,404	2,602	2,044	254
(76.5, 85.0]	79	4,472	1,977	1,353	115

Cross Tabulation (2/3)

Variable	Insurance Coverage Type (GRP, DIR, PUB)				
	NNN	NNY	NY_	Y1Y	YNN
PEMLR: Major labor force recode					
Universe: All persons					
NIU	2,849	10,183	1,494	1,161	17,734
Employed - at work	7,178	6,826	5,136	3,181	46,957
Employed - absent	328	471	335	161	1,914
Unemployed - on layoff	252	341	136	72	797
Unemployed - looking	479	641	138	121	909
Not in labor force - retired	543	11,004	5,087	3,754	1,768
Not in labor force - disabled	437	4,110	405	359	732
Not in labor force - other	2,969	5,169	1,510	686	9,354

Cross Tabulation (3/3)

Variable	Insurance Coverage Type (GRP, DIR, PUB)				
	NNN	NNY	NY_	Y1Y	YNN
SS_YN: Who received social security payments either for themselves or as combined payments with other family members ? Universe: All persons aged 15+					
Niu	2,431	10,167	1,488	1,160	17,629
Yes	397	13,477	5,642	4,471	1,228
No	12,207	15,101	7,111	3,864	61,308

Data Encoding

- The input dataset is encoded in the correct format (by instantiating a class)
  - Continuous NIU (not in universe) value $\rightarrow$ 0
  - Categorical value $\rightarrow$ 0 up to a positive integer
- The state of an instance is maintained by two user-defined attached atttributes
  - `dataset`: a pandas DataFrame extended by the `data` accessor
  - `metadata`: a Python list
- The user-defined nonstatic methods `encodecat` and `encodecont` for encoding categorical and continuous features change the object into multiple states
- This dissertation excessively uses the shallow copies of attributes by calling the method `copy` to protect the originals
  - Unlike a deep copy, a shallow copy inserts reference to an original object to the extent possible

Sampling using SelectKBest

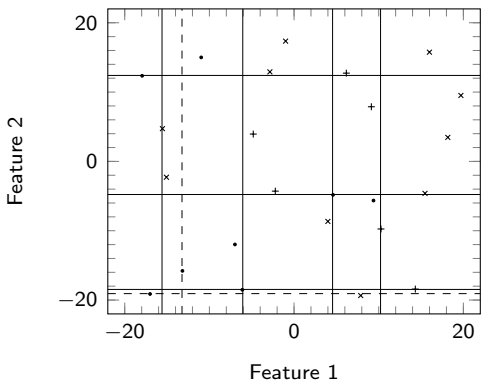
- Because the proposed classifier is exponentially expensive, certain features are preselected by evaluating their scores against a target variable
- In addition, 100 out of 157,681 survey participants are sampled of equal class size by calling two built-in methods `groupby` and `sample`
- Due to its random nature, the sampling result changes in each call
- The use of the model is illustrated with only three preselected features based on three highest scores based on the mutual information for a discrete target

Setting Number of Variable Splits

- An appropriate number of splits on an individual feature in the health insurance dataset of all noninfant survey participants
- In the case of three splits, up to two splits are allowed on the feature SS_YN representing the answer, including NIU (not in universe), to the yes/no question regarding social security payment
- The column of these numbers is inserted into the DataFrame as an additional information directly through the user-defined pandas accessor `info` without explicit class instantiation

Proposed Classifier

Splitting Values for Three-Class Dataset (N = 24)



Type of features: continuous
Hyperparameters: $M = 500 \mid m_j = 0.05$
Feature 1: $5 \rightarrow 4$ splitting values

- x _ -15.6 _ -13.23 _ -6.06 _ 4.55 _ 10.22 _ x
- _ -15.6 _ -6.06 _ 4.55 _ 10.22 _

Feature 2: $5 \rightarrow 3$ splitting values

- x _ -19.1 _ -18.5 _ -4.8 _ 12.4 _ 499.95 _ x
- _ -18.5 _ -4.8 _ 12.4 _

Training accuracy: 100%
MILP solver: CPLEX
Execution time: 58 seconds

• class 0 × class 1 + class 2 —— minimal splitting value - - - redundant splitting value

Training Data

Setup

- Each instance $\tilde{x}^i = (\tilde{x}^i_{\tilde{j}})_{1 \leq \tilde{j} \leq \tilde{d}} \in \mathbb{R}^{\tilde{d}}$ where $1 \leq i \leq N$
- Class label: $0 \leq k \leq n$
- $\tilde{d}$ features/factors $(\tilde{j}) \implies$ up to d contributing factors (j) with $d \leq \tilde{d}$
- Usually, $d << \tilde{d}$
- Reduced instance: $x^i = (x^i_j)_{1 \leq j \leq d} \in \mathbb{R}^d$ (of less dimension)
- y^i_k indicates whether x^i is in class k

Linear Constraints for Selecting

$$x^i_j = \sum_{j=1}^{d} c_{j,\tilde{j}} \tilde{x}^i_{\tilde{j}}$$

$$\sum_{\tilde{j}=1}^{\tilde{d}} c_{j,\tilde{j}} \leq 1$$

$$\sum_{j=1}^{d} c_{j,\tilde{j}} \leq 1$$

$$c_{j,\tilde{j}} \in \{0, 1\}$$

Feature Selection

Criteria for Selecting Features

An original feature $1 \leq \tilde{j} \leq \tilde{d}$ is selected and considered significant only when

$$\sum_{j=1}^{d} c_{j,\tilde{j}} = 1$$

and it becomes a new feature $1 \leq j \leq d << \tilde{d}$, uniquely, for $c_{j,\tilde{j}} = 1$

Previous Linear Constraints

$$x_j^i = \sum_{j=1}^{d} c_{j,\tilde{j}} \tilde{x}_{\tilde{j}}^i$$

$$\sum_{\tilde{j}=1}^{\tilde{d}} c_{j,\tilde{j}} \leq 1$$

$$\sum_{j=1}^{d} c_{j,\tilde{j}} \leq 1$$

$$c_{j,\tilde{j}} \in \{0, 1\}$$

Splitting Values (1/2)

Splitting Values

- New feature $1 \leq j \leq d$ has $p_j \geq 0$ splitting values: $b_{j,1} \leq \ldots \leq b_{j,p_j}$
- Two endpoints (subsequently treated as decision variables) are assumed: $b_{j,0} = -M$ and $b_{j,p_j+1} = M$ for sufficiently large positive M such as $\max\{|x_j^i|\}$
- $\alpha_{j,q}^i = I(x_j^i \in (b_{j,q}, b_{j,q+1}))$
- $[b_{j,q} + m_j, b_{j,q+1} - m_j]$ for sufficiently small m_j such as $m_j = \frac{1}{2}\min\{|x_j^{i_1} - x_j^{i_2}| : x_j^{i_1} \neq x_j^{i_2}\}$

Component j of Box: $[b_{j,q}, b_{j,q+1}]$

Nonlinear Constraints

$$x_j^i \in \sum_{q=0}^{p_j} \alpha_{j,q}^i [b_{j,q} + m_j, b_{j,q+1} - m_j]$$

$$= \sum_{q=0}^{p_j} [l_{j,q}^i, r_{j,q}^i]$$

$$\sum_{q=0}^{p_j} \alpha_{j,q}^i = 1$$

$$\alpha_{j,q}^i \in \{0, 1\}$$

Splitting Values (2/2)

Previous Nonlinear Contraints

$$x_j^i = \sum_{q=0}^{p_j} [l_{j,q}^i, r_{j,q}^i]$$
$$l_{j,q}^i = \alpha_{j,q}^i (b_{j,q} + m_j)$$
$$r_{j,q}^i = \alpha_{j,q}^i (b_{j,q+1} - m_j)$$
$$\sum_{q=0}^{p_j} \alpha_{j,q}^i = 1,\ \alpha_{j,q}^i \in \{0,1\}$$

Recall that

$$\begin{aligned}\alpha_{j,q}^i &= I(x_j^i \in (b_{j,q}, b_{j,q+1})) \\ &= I(x_j^i \in [b_{j,q} + m_j, b_{j,q+1} - m_j])\end{aligned}$$

Equivalent Linear Constraints

$$x_j^i = \sum_{q=0}^{p_j} [l_{j,q}^i, r_{j,q}^i]$$
$$l_{j,q}^i \in [-M, b_{j,q} + m_j] + M(1 - \alpha_{j,q}^i)$$
$$l_{j,q}^i \in [b_{j,q} + m_j, M] - M(1 - \alpha_{j,q}^i)$$
$$r_{j,q}^i \in [-M, b_{j,q+1} - m_j] + M(1 - \alpha_{j,q}^i)$$
$$r_{j,q}^i \in [b_{j,q+1} - m_j, M] - M(1 - \alpha_{j,q}^i)$$
$$\sum_{q=0}^{p_j} \alpha_{j,q}^i = 1,\ \alpha_{j,q}^i \in \{0,1\}$$

Decision Boxes (1/2)

Decision Box β

$$S_\beta = \prod_{j=1}^{d} \sum_{q=0}^{p_j} \beta_{j,q}[b_{j,q}, b_{j,q+1}]$$

where

- $\beta_{j,q} = I(\text{Entry } j = [b_{j,q}, b_{j,q+1}])$
- small $b_{j,0} < 0$ and large $b_{j,p_j+1} > 0$
- $\sum_{q=0}^{p_j} \beta_{j,q} = 1$
- $\beta = \sum_{j=1}^{d} \left[\prod_{j_0=0}^{j-1} (p_{j_0} + 1)\right] \left[\sum_{q=0}^{p_j} q\beta_{j,q}\right]$ with $p_0 = 0$

Locations of Instances

An instance $x^i \in \mathbb{R}^d$ is located in one of the $B = (p_1 + 1) \cdots (p_d + 1)$ decision boxes labeled by $0 \le \beta \le B - 1$

$$\sum_{j=1}^{d} \left[\prod_{j_0=0}^{j-1} (p_{j_0} + 1)\right] \left[\sum_{q=0}^{p_j} q\alpha_{j,q}^i\right] = \sum_{\beta=0}^{B-1} \beta\gamma_\beta^i$$

$$\sum_{\beta=0}^{B-1} \gamma_\beta^i = 1$$

$$\gamma_\beta^i \in \{0, 1\}$$

$$\boxed{\alpha_{j,q}^i = I(x_j^i \in [b_{j,q} + m_j, b_{j,q+1} - m_j])}$$

Decision Boxes (2/2)

Predicted Class Labels of Boxes

By majority voting, a decision box β therefore predicts exactly one class label from the following set

$$\Theta_\beta = \underset{0 \le k \le n}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} y_k^i \gamma_\beta^i \right\}$$

where

- $y_k^i = I(x^i \in \text{Class } k)$
- $\gamma_\beta^i = I(x^i \in \text{Box } \beta)$
- $\alpha_{j,q}^i = I(x_j^i \in [b_{j,q} + m_j, b_{j,q+1} - m_j])$

Locations of Instances (Revisited)

An instance $x^i \in \mathbb{R}^d$ is located in one of the $B = (p_1 + 1) \cdots (p_d + 1)$ decision boxes labeled by $0 \le \beta \le B - 1$

$$\sum_{j=1}^{d} \left[\prod_{j_0=0}^{j-1} (p_{j_0} + 1) \right] \left[\sum_{q=0}^{p_j} q \alpha_{j,q}^i \right] = \sum_{\beta=0}^{B-1} \beta \gamma_\beta^i$$

$$\sum_{\beta=0}^{B-1} \gamma_\beta^i = 1$$

$$\gamma_\beta^i \in \{0, 1\}$$

Misclassification

Number of Misclassified Instances

$$\Theta_\beta = \underset{0\le k\le n}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} y_k^i \gamma_\beta^i \right\}$$

The number of misclassified instances is

$$N - \sum_{\beta=0}^{B-1} \max_{0\le k\le n} \left\{ \sum_{i=1}^{N} y_k^i \gamma_\beta^i \right\}$$

$$N + \sum_{\beta=0}^{B-1} \min_{0\le k\le n} \left\{ -\sum_{i=1}^{N} y_k^i \gamma_\beta^i \right\}$$

Linear Program

To compute $\min_{0\le k\le n} \left\{ -\sum_{i=1}^{N} y_k^i \gamma_\beta^i \right\}$:

minimize h_β

subject to

$$h_\beta \ge -\sum_{i=1}^{N} y_k^i \gamma_\beta^i - N z_{\beta,k}$$

$$\sum_{k=0}^{n} z_{\beta,k} = n$$

$$z_{\beta,k} \in \{0, 1\}$$

Linearization of Misclassification

Linear Program ($\mathcal{P}$)

minimize h_β

subject to

$$h_\beta \geq -\sum_{i=1}^{N} y_k^i \gamma_\beta^i - N z_{\beta,k}$$

$$\sum_{k=0}^{n} z_{\beta,k} = n$$

$$z_{\beta,k} \in \{0,1\}.$$

Optimal Objective Value

Subproblem $\mathcal{P}_{k_0}$: $z_{\beta,k} = 0 \iff k = k_0$

$$h_\beta \geq -\sum_{i=1}^{N} y_{k_0}^i \gamma_\beta^i = 0 - \sum_{i=1}^{N} y_{k_0}^i \gamma_\beta^i$$

$$\geq -\sum_{i=1}^{N} y_k^i \gamma_\beta^i - N z_{\beta,k}$$

$$\boxed{\min(\mathcal{P}_{k_0}) = -\sum_{i=1}^{N} y_{k_0}^i \gamma_\beta^i}$$

Selection Model for Continuous Features (1/3)

$$\text{minimize} \quad \sum_{\beta=0}^{B-1} h_\beta$$

$$\text{subject to} \quad x_j^i - \sum_{j=1}^{d} \tilde{x}_j^i c_{j,\tilde{j}} = 0$$

$$\sum_{\tilde{j}=1}^{\tilde{d}} c_{j,\tilde{j}} \leq 1$$

$$\sum_{j=1}^{d} c_{j,\tilde{j}} \leq 1$$

$$b_{j,q+1} - b_{j,q} \geq 0$$

$$\ldots$$

$$\text{minimize} \quad \sum_{\beta=0}^{B-1} h_\beta$$

$$\text{subject to} \quad \ldots$$

$$\sum_{j=1}^{d} \left[\prod_{j_0=0}^{j-1} (p_{j_0} + 1) \right] \left[\sum_{q=0}^{p_j} q \alpha_{j,q}^i \right] - \sum_{\beta=0}^{B-1} \beta \gamma_\beta^i = 0$$

$$\sum_{q=0}^{p_j} \alpha_{j,q}^i = 1$$

$$\sum_{\beta=0}^{B-1} \gamma_\beta^i = 1$$

$$\ldots$$

Selection Model for Continuous Features (2/3)

$$\text{minimize} \quad \sum_{\beta=0}^{B-1} h_\beta$$

$$\text{subject to} \quad \ldots$$

$$\left(\cancel{x_j^i \leq} \sum_{j=1}^{d} \tilde{x}_j^i c_{j,\tilde{j}}\right) - \sum_{q=0}^{p_j} l_{j,q}^i \geq 0$$

$$l_{j,q}^i + M\alpha_{j,q}^i \geq 0$$

$$l_{j,q}^i - M\alpha_{j,q}^i \leq 0$$

$$l_{j,q}^i - b_{j,q} + M\alpha_{j,q}^i \leq M + m_j$$

$$l_{j,q}^i - b_{j,q} - M\alpha_{j,q}^i \geq -M + m_j$$

$$\ldots$$

$$\text{minimize} \quad \sum_{\beta=0}^{B-1} h_\beta$$

$$\text{subject to} \quad \ldots$$

$$\left(\cancel{x_j^i \leq} \sum_{j=1}^{d} \tilde{x}_j^i c_{j,\tilde{j}}\right) - \sum_{q=0}^{p_j} r_{j,q}^i \leq 0$$

$$r_{j,q}^i + M\alpha_{j,q}^i \geq 0$$

$$r_{j,q}^i - M\alpha_{j,q}^i \leq 0$$

$$r_{j,q}^i - b_{j,q+1} + M\alpha_{j,q}^i \leq M - m_j$$

$$r_{j,q}^i - b_{j,q+1} - M\alpha_{j,q}^i \geq -M - m_j$$

$$\ldots$$

Selection Model for Continuous Features (3/3)

$$\text{minimize} \quad \sum_{\beta=0}^{B-1} h_\beta$$

$$\text{subject to} \quad \ldots$$

$$h_\beta + \sum_{i=1}^{N} y_k^i \gamma_\beta^i + N z_{\beta,k} \geq 0$$

$$\sum_{k=0}^{n} z_{\beta,k} = n$$

$$\cancel{x_i}, l_{j,q}^i, r_{j,q}^i, b_{j,q}, h_\beta \in \mathbb{R}$$

$$c_{j,\tilde{j}}, \alpha_{j,q}^i, \gamma_\beta^i, z_{\beta,k} \in \{0,1\}$$

The 0-1 MILP problem provides a training accuracy of

$$1 + \frac{\sum_{\beta=0}^{B-1} h_\beta^*}{N} \leq 1$$

Selection of Continuous Features in Mixed Data

Mixed-Type Features

- $\tilde{x}_{\tilde{j}}^{i} \in \mathbb{R}$: continuous/categorical
- Index sets:
  $\tilde{\mathcal{C}}_{\text{cont}} \cup \tilde{\mathcal{C}}_{\text{cat}} = \{1, 2, \ldots, \tilde{d}\}$

Selection of Continuous Features

- New index sets: $\mathcal{C}_{\text{cont}}$ and $\mathcal{C}_{\text{cat}}$

$$|\mathcal{C}_{\text{cont}}| \leq |\tilde{\mathcal{C}}_{\text{cont}}| \Leftarrow \mathcal{C}_{\text{cont}} \subseteq \tilde{\mathcal{C}}_{\text{cont}}$$
$$|\mathcal{C}_{\text{cat}}| = |\tilde{\mathcal{C}}_{\text{cat}}| \Leftarrow \mathcal{C}_{\text{cat}} = \tilde{\mathcal{C}}_{\text{cat}}$$
$$\mathcal{C}_{\text{cont}} \cup \mathcal{C}_{\text{cat}} = \{1, 2, \ldots, d\}$$

Continuous Data Type

$$x_j^i = \sum_{\tilde{j} \in \tilde{\mathcal{C}}_{\text{cont}}} c_{j,\tilde{j}} \tilde{x}_{\tilde{j}}^i, \quad j \in \mathcal{C}_{\text{cont}}$$

$$\sum_{\tilde{j} \in \tilde{\mathcal{C}}_{\text{cont}}} c_{j,\tilde{j}} \leq 1, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$\sum_{j \in \mathcal{C}_{\text{cont}}} c_{j,\tilde{j}} \leq 1, \qquad \tilde{j} \in \tilde{\mathcal{C}}_{\text{cont}}$$

$$c_{j,\tilde{j}} \in \{0, 1\}, \qquad (j, \tilde{j}) \in \mathcal{C}_{\text{cont}} \times \mathcal{C}_{\text{cont}}$$

$$\boxed{\sum_{(j,\tilde{j}) \in \mathcal{C}_{\text{cont}} \times \tilde{\mathcal{C}}_{\text{cont}}} c_{j,\tilde{j}} \leq |\mathcal{C}_{\text{cont}}|}$$

Classification of Mixed Data

A selected, rearranged component $x_j^i \in \mathbb{R}$ is now either continuous or categorical

- A continuous feature $j \in \mathcal{C}_{\text{cont}}$ is similarly assumed to have p_j splitting points, namely $b_{j,q} \in \mathbb{R}$ where $1 \leq q \leq p_j$
- A categorical feature $j \in \mathcal{C}_{\text{cat}}$ comprises finite discrete values which are also assumed to form $p_j + 1$ new small groups labeled with $0 \leq u_j \leq p_j$

Decision Boxes for Mixed Data

A box $0 \leq \beta \leq B-1$ is identified by binary $(\beta_{j,q})_{j \in \mathcal{C}_{\text{cont}}}$ and integer $(u_j)_{j \in \mathcal{C}_{\text{cat}}}$

$$S_\beta = \prod_{j \in \mathcal{C}_{\text{cont}}} \sum_{q=0}^{p_j} \beta_{j,q} [b_{j,q}, b_{j,q+1}] \times \prod_{j \in \mathcal{C}_{\text{cat}}} \{u_j\}$$

$$\beta = \sum_{j \in \mathcal{C}_{\text{cont}}} \left[\prod_{0 \leq j_0 < j} (p_{j_0} + 1) \right] \left[\sum_{q=0}^{p_j} q \beta_{j,q} \right] + \sum_{j \in \mathcal{C}_{\text{cat}}} \left[\prod_{0 \leq j_0 < j} (p_{j_0} + 1) \right] u_j$$

$$\sum_{q=0}^{p_j} \beta_{j,q} = 1, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$\beta_{j,q} \in \{0,1\}, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$u_j \in \{0,1,\ldots,p_j\}, \qquad j \in \mathcal{C}_{\text{cat}}$$

Reassignment of Categorical Labels

For a categorical feature $j \in \mathcal{C}_{\text{cat}}$, an original categorical label $x_j^i \in \mathbb{R}$ is reassigned to a new integer group label $0 \leq v_{j,x_j^i} \leq p_j$

$$\sum_{\beta=0}^{B-1} \beta\gamma_\beta^i = \sum_{j\in\mathcal{C}_{\text{cont}}} \left[\prod_{0\leq j_0<j} (p_{j_0}+1)\right] \left[\sum_{q=0}^{p_j} q\alpha_{j,q}^i\right] + \sum_{j\in\mathcal{C}_{\text{cat}}} \left[\prod_{0\leq j_0<j} (p_{j_0}+1)\right] v_{j,x_j^i}$$

$$\sum_{q=0}^{p_j} \alpha_{j,q}^i = 1, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$\sum_{\beta=0}^{B-1} \gamma_\beta^i = 1$$

$$\beta_{j,q} \in \{0,1\}, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$v_{j,x_j^i} \in \{0,1,\ldots,p_j\}, \qquad j \in \mathcal{C}_{\text{cat}}$$

Feature Selection

Categorical Features

- $f_j \in \{0,1\}$ = whether a categorical feature j is selected
- All categorical labels of an insignificant feature are grouped together
- Necessary condition: $-Mf_j \leq v_{j,x_j^i} \leq Mf_j$
- If at most d_{cat} out of $|\mathcal{C}_{\text{cat}}|$ categorical features are of interest:

$$\sum_{j \in \mathcal{C}_{\text{cat}}} f_j \leq d_{\text{cat}}.$$

Mixed-Type Features

- At most $|\mathcal{C}_{\text{cont}}| + d_{\text{cat}} \leq d \leq \tilde{d}$ contributing factors
  - $|\mathcal{C}_{\text{cont}}| \leq |\tilde{\mathcal{C}}_{\text{cont}}|$ continuous
  - $d_{\text{cat}} \leq |\mathcal{C}_{\text{cat}}| = |\tilde{\mathcal{C}}_{\text{cat}}|$ categorical

$$\sum_{(j,\tilde{j}) \in \mathcal{C}_{\text{cont}} \times \tilde{\mathcal{C}}_{\text{cont}}} c_{j,\tilde{j}} + \sum_{j \in \mathcal{C}_{\text{cat}}} f_j \leq d.$$

Investigation of Contributing Factors

- $\tilde{j} \in \tilde{\mathcal{C}}_{\text{cont}}$: $c_{j,\tilde{j}} = 1$ for some $j \in \mathcal{C}_{\text{cont}}$
- $\tilde{j} \in \tilde{\mathcal{C}}_{\text{cat}} \leftrightarrow j \in \mathcal{C}_{\text{cat}}$: v_{j,x_j^i} nonconstant across all x^i

Final Selection Model (1/5)

$$\text{minimize} \quad \sum_{\beta=0}^{B-1} h_\beta$$

$$\text{subject to}$$

$$\sum_{\tilde{j} \in \tilde{\mathcal{C}}_{\text{cont}}} c_{j,\tilde{j}} \leq 1, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$\sum_{j \in \mathcal{C}_{\text{cont}}} c_{j,\tilde{j}} \leq 1, \qquad j \in \tilde{\mathcal{C}}_{\text{cont}}$$

$$b_{j,q+1} - b_{j,q} \geq 0, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$\sum_{\tilde{j} \in \tilde{\mathcal{C}}_{\text{cont}}} \tilde{x}^i_{\tilde{j}} c_{j,\tilde{j}} - \sum_{q=0}^{p_j} l^i_{j,q} \geq 0, \qquad j \in \mathcal{C}_{\text{cont}}$$

Final Selection Model (2/5)

$$\sum_{\tilde{j}\in\tilde{\mathcal{C}}_{\text{cont}}} \tilde{x}^i_{\tilde{j}} c_{j,\tilde{j}} - \sum_{q=0}^{p_j} r^i_{j,q} \leq 0, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$l^i_{j,q} + M\alpha^i_{j,q} \geq 0, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$l^i_{j,q} - M\alpha^i_{j,q} \leq 0, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$l^i_{j,q} - b_{j,q} + M\alpha^i_{j,q} \leq M + m_j, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$l^i_{j,q} - b_{j,q} - M\alpha^i_{j,q} \geq -M + m_j, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$r^i_{j,q} + M\alpha^i_{j,q} \geq 0, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$r^i_{j,q} - M\alpha^i_{j,q} \leq 0, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$r^i_{j,q} - b_{j,q+1} + M\alpha^i_{j,q} \leq M - m_j, \qquad j \in \mathcal{C}_{\text{cont}}$$

$$r^i_{j,q} - b_{j,q+1} - M\alpha^i_{j,q} \geq -M - m_j, \qquad j \in \mathcal{C}_{\text{cont}}$$

Final Selection Model (3/5)

$$\sum_{j \in \mathcal{C}_{\text{cont}}} \left[\prod_{0 \le j_0 < j} (p_{j_0} + 1) \right] \left[\sum_{q=0}^{p_j} q \alpha_{j,q}^{i} \right] + \sum_{j \in \mathcal{C}_{\text{cat}}} \left[\prod_{0 \le j_0 < j} (p_{j_0} + 1) \right] v_{j,x_j^i} - \sum_{\beta=0}^{B-1} \beta \gamma_{\beta}^{i} = 0$$

$$\sum_{q=0}^{p_j} \alpha_{j,q}^{i} = 1, \qquad j \in \mathcal{C}_{\text{cont}}$$

Final Selection Model (4/5)

$$v_{j,x_j^i} + Mf_j \geq 0, \qquad j \in \mathcal{C}_{\text{cat}}$$

$$v_{j,x_j^i} - Mf_j \leq 0, \qquad j \in \mathcal{C}_{\text{cat}}$$

$$\sum_{(j,\tilde{j}) \in \mathcal{C}_{\text{cont}} \times \tilde{\mathcal{C}}_{\text{cont}}} c_{j,\tilde{j}} + \sum_{j \in \mathcal{C}_{\text{cat}}} f_j \leq d$$

$$\sum_{\beta=0}^{B-1} \gamma_\beta^i = 1$$

$$h_\beta + \sum_{i=1}^{N} y_k^i \gamma_\beta^i + N z_{\beta,k} \geq 0$$

$$\sum_{k=0}^{n} z_{\beta,k} = n$$

Final Selection Model (5/5)

$$
\begin{aligned}
l^i_{j,q}, r^i_{j,q}, b_{j,q} &\in \mathbb{R}, && j \in \mathcal{C}_{\text{cont}} \\
h_\beta &\in \mathbb{R} && \\
c_{j,\tilde{j}} &\in \{0,1\}, && (j,\tilde{j}) \in \mathcal{C}_{\text{cont}} \times \tilde{\mathcal{C}}_{\text{cont}} \\
\alpha^i_{j,q} &\in \{0,1\}, && j \in \mathcal{C}_{\text{cont}} \\
f_j &\in \{0,1\}, && j \in \mathcal{C}_{\text{cat}} \\
v_{j,x^i_j} &\in \{0,1,\ldots,p_j\}, && j \in \mathcal{C}_{\text{cat}} \\
\alpha^i_{j,q}, \gamma^i_\beta, z_{\beta,k} &\in \{0,1\} &&
\end{aligned}
$$

CPLEX OPL Modeling (1/2)

- CPLEX optimizer (version 22.1.1) uses in-memory computation
- Although achieving higher performance, manual adjustment of internal optimization procedures such as a node selection during branching and a combination of multiple techniques in cut generation is beyond the scope of this dissertation
- Huge tree data structure as a result of large MILP
- Multiple lock-free nodes can be executed in parallel by utilizing all CPU cores
- As more solutions are explored, the branch-and-cut tree grows larger
  - When its size exceeds its upper limit, which is set at 10^{75} MB by default, the optimization process terminates
  - The solver also stops when a memory is exhausted or a disk is fully occupied depending on whether node files are stored in memory or on disk

CPLEX OPL Modeling (2/2)

- Classification files are written in Optimization Programming Language (OPL)
- The cardinality of a new continuous component $|\mathcal{C}_{\text{cont}}|$ is assumed to be the minimum of its given counterpart $|\tilde{\mathcal{C}}_{\text{cont}}|$ and an upper bound on the number of significant features d
  - A new continuous feature with splitting values may turn unselected
- Setup
  - The sufficiently small positive number m_0 is set to be 0.01
  - The execution time is limited up to 24 hours or one day
  - Every MIP solution, feasible but not necessarily optimal, is recorded, thereby calling a CPLEX solver multiple times
  - After the working memory exceeds 2 GB, some nodes are transferred to disk in compressed form
  - The uncompressed tree size is limited to 200 GB

CPLEX Parameters (1/3)

Parameter	Description	Default	Chosen
`cplex.intsollim`	MIP solution number limit		1
`cplex.tilim`	Time limit per optimizer call (in seconds)		1 day
`cplex.threads`	Parallel threads	0 (≤ 32)	
`cplex.workmem`	Working memory before compression and swap (in MB)	2048	
`cplex.trelim`	Uncompressed tree limit (in MB)	10^{75}	200 GB

CPLEX Parameters (2/3)

Parameter	Description
`cplex.nodefileind`	Node storage file switch 0: No node file 1: Node file in memory and compressed (default) 2: Node file on disk 3: Node file on disk and compressed (chosen)

CPLEX Parameters (3/3)

Parameter	Description
`cplex.status`	Solution status code 1: Optimal for simplex and barrier methods 11: Time limit exceeded 101: Optimal for MIP model 102: Optimal within predefined MIP gap tolerance 104: Limit on mixed integer solutions 111: Tree memory limit exceeded and integer solution found 112: Tree memory limit exceeded and no integer solution

Recalculation of Decision Boxes (1/3)

- Some of d selected features may be trivial (not significant)
- Consider the optimal splitting values $b^*_{j,q}$ and v^*_{j,x_j^i}
- New continuous feature $j \in \mathcal{C}_{\text{cont}} \rightarrow$ Examine $b^*_{j,q}$
  - Two consecutive splitting values covers an entire dataset
  - Turn unselected: $c^*_{j,\tilde{j}} = 0$ for all $\tilde{j} \in \tilde{\mathcal{C}}_{\text{cont}}$
- New categorical feature $j \in \mathcal{C}_{\text{cat}} \rightarrow$ Examine v^*_{j,x_j^i}
  - All categorical values are reallocated to the same group
- This leads to excessive number of decision boxes
- Determine which two distinct boxes can be merged
  - All numerical decision box labels are recalculated through a transformation g to new labels in a final feature space

Recalculation of Decision Boxes (2/3)

Suppose only d' out of d features are finally selected. The feature map $\sigma : \{0, 1, \ldots, d\} \to \{-1\} \cup \{0, 1, \ldots, d'\}$ is defined by

$$\sigma(j) = \begin{cases} \text{feature in new space}, & \text{for finally selected feature } j \\ -1, & \text{for finally unselected feature } j \\ 0, & \text{if } j = 0. \end{cases}$$

There is a one-to-one corresponding between j and $\sigma(j) \geq 0$, and the image of σ includes $0, 1, \ldots, d'$. Consider a decision box $1 \leq \beta \leq B$. Define its position along a feature j by

$$q_j = \begin{cases} \sum_{q=0}^{p_j} q\beta_{j,q}, & \text{for continuous feature } j \\ u_j, & \text{for categorical feature } j. \end{cases}$$

Recalculation of Decision Boxes (3/3)

Theorem (Box Merging)

A transformation g maps an existing box label β to a new label in a final feature space:

$$g(\beta) - g(\beta - 1) = -\sum_{j=1}^{w-1} p_j \prod_{j' \in \Sigma_j} (p_{j'} + 1) + 1 \cdot \prod_{j' \in \Sigma_w} (p_{j'} + 1)$$

where $g(0) = 0$ and $\Sigma_j = \{j' : 0 \leq \sigma(j') < \sigma(j)\}$.

Proof.

Let $w = \min\{j : q_j \neq 0\}$. If $w = 1$, then both positions of the current box β and the previous counterpart $\beta - 1$ along the first feature differ by 1. For $w > 1$, the previous box $\beta - 1$ locates at position p_j along every feature $j < w$, and the position of both boxes at feature w differs by 1. □

Results on Health Insurance

Training Data (1/4)

- Sample size of 100 (25 per class)
- Three preselected features: A_AGE, PEMLR and SS_YN
- Although survey participants are unique, some sample records can be the same in feature and even in target due to initial preselection of features and resultant partial loss of personal information
- Two contributing factors out of three are investigated based solely on highest training accuracy

Training Data (2/4)

Preselected Variable	Insurance Coverage Type (GRP, DIR, PUB)				
	NNN	NNY	NY_	Y1Y	YNN
A_AGE: Age					
Universe: All persons					
(1.917, 18.6]	4	8	2	0	5
(18.6, 35.2]	10	2	1	4	8
(35.2, 51.8]	5	1	5	2	5
(51.8, 68.4]	1	4	8	6	2
(68.4, 85.0]	0	5	4	8	0

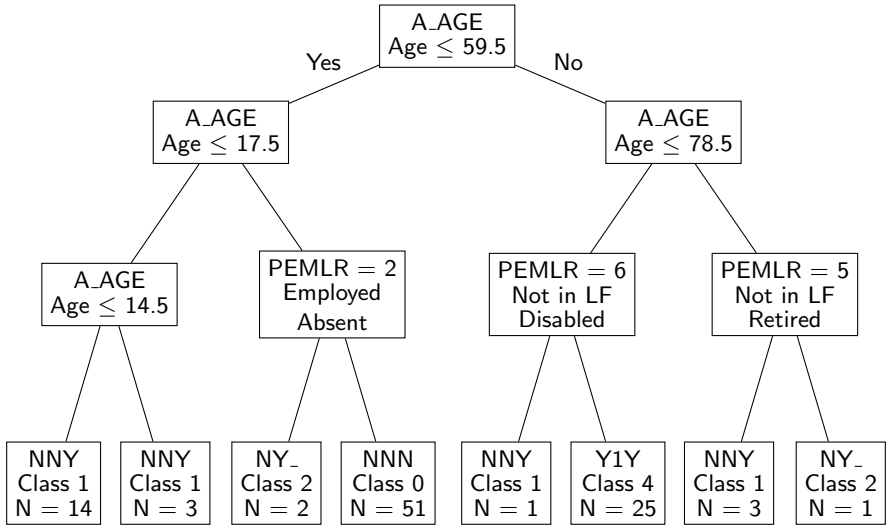
Training Data (3/4)

Preselected Variable	Insurance Coverage Type (GRP, DIR, PUB)				
	NNN	NNY	NY_	Y1Y	YNN
PEMLR: Major labor force recode					
Universe: All persons					
0: NIU	4	5	2	0	4
1: Employed - at work	8	3	7	9	12
2: Employed - absent	0	0	3	1	0
3: Unemployed - on layoff	1	1	0	0	0
4: Unemployed - looking	1	1	1	0	2
5: Not in labor force - retired	0	5	5	9	0
6: Not in labor force - disabled	0	2	1	0	0
7: Not in labor force - other	6	3	1	1	2

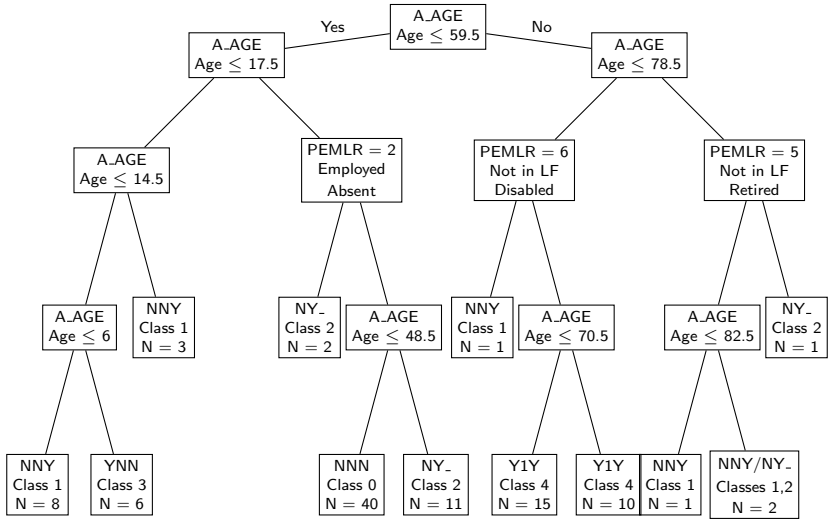
Training Data (4/4)

Preselected Variable	Insurance Coverage Type (GRP, DIR, PUB)				
	NNN	NNY	NY_	Y1Y	YNN
SS_YN: Who received social security payments either for themselves or as combined payments with other family members ? Universe: All persons aged 15+					
0: NIU	3	5	2	0	4
1: Yes	0	9	7	10	1
2: No	17	6	11	10	15

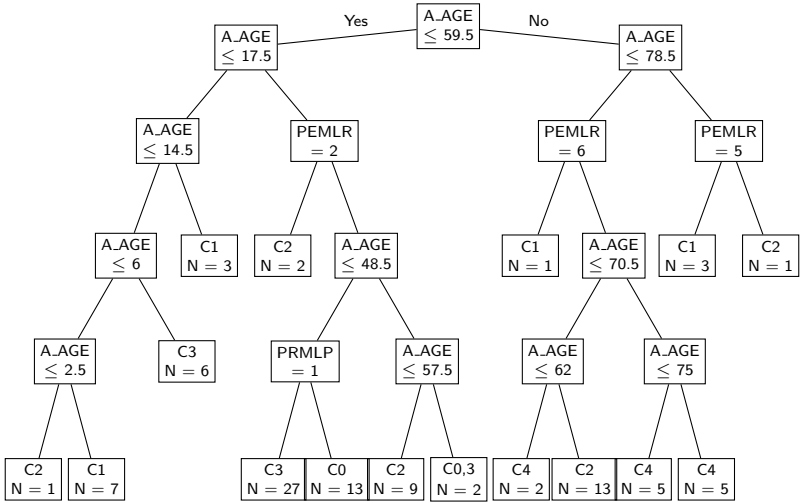
Decision Tree (Depth 3 & Accuracy 45%)



Decision Tree (Depth 4 & Accuracy 50%)



Decision Tree (Depth 5 & Accuracy 54%)



Comparison: Decision Tree vs. Proposed Model

Classification Model		Num of Splitting Values				Boxes	Accuracy (%)	Time (min)
Model	Spec	A_AGE	PEMLR	SS_YN	Total			
Tree	Dep 3	4	3	0	7	8	45	0.08
	Dep 4	8	3	0	11	12	50	
	Dep 5	12	3	0	15	16	54	
Proposed	Iter 13	3	3	0	6	16	51	78.88
	Iter 14	3	3	0	6	16	51	82.02
	Iter 15	3	3	0	6	16	51	209.93

Splitting Values

Classification Model		Splitting Values			Accuracy (%)
Model	Spec	A_AGE	PEMLR	SS_YN	
Tree	Dep 3	14.5, 17.5, 59.5, 78.5	{2}, {5}, {6}, {0, 1, 3, 4, 7}	–	45
	Dep 4	6, 14.5, 17.5, 48.5, 59.5, 70.5, 78.5, 82.5	{2}, {5}, {6}, {0, 1, 3, 4, 7}	–	50
	Dep 5	2.5, 6, 14.5, 17.5, 48.5, 57.5, 59.5, 62, 70.5, 75, 78.5	{2}, {5}, {6}, {0, 1, 3, 4, 7}	–	54
Proposed	Iter 13	24.99, 55.99, 64.99	{2}, {1}, {3, 4, 5, 7}, {0, 6}	–	51
	Iter 14 to 15	24.01, 55.99, 64.99	{2}, {1}, {3, 4, 5, 7}, {0, 6}	–	51

Performance

Iter	Accuracy (%)			Time (min)		Min Storage (GB)			Rel Gap	Inconsistent
	True	CPLEX	Reported	Each	Accum	Tree	Nodes	Comp		
1			20	0	0				279	
2	38	35	28	0.03	0.03				27.57	41
3	38	35	31	0.01	0.04				22.14	41
4	38	35	36	0.01	0.06				17.25	41
5	38	35	38	0.03	0.09				15.5	41
6	40	36	39	13.3	13.39	0.99	0	0	8.67	41
7	40	30	40	5.27	18.66	1.24	0	0	8.42	100
8	43	40	43	4.64	23.3	2.74	0.49	0.45	7.75	41
9	44	42	44	7.67	30.97	3.68	1.3	1.18	7.54	41
10	47	47	46	37.23	68.2	3.35	1.34	1.19	7.01	
11	48	48	48	1.18	69.38	3.46	1.5	1.32	6.67	
12	50	50	49	7.17	76.55	4.11	1.64	1.45	6.51	
13	51	51	50	2.33	78.88	8.13	5.92	5.17	6.35	
14	51	51	51	3.14	82.02	9.06	7	6.13	6.2	
15	51	51	51	127.91	209.93	192.68	190.58	167.06	6.08	

Feature Selection (Iteration 15)

Selected Variable			Group	Member	
Ind	Symbol	Type		Index	Label
1	A_AGE	Cont	0	$(-\infty, 24.01)$	Below 24
			1	$(24.01, 55.99)$	Between 25 and 55
			2	$(55.99, 64.99)$	Between 56 and 64
			3	$(64.99, \infty)$	Above 65
2	PEMLR	Cat	0	2	Employed - absent
			1	1	Employed - at work
			2	3	Unemployed - on layoff
				4	Unemployed - looking
				5	Not in labor force - retired
				7	Not in labor force - other
			3	0	NIU (not in universe)
				6	Not in labor force - disabled

Decision Regions (Iteration 15)

Decision Region			Predicted Classes		Num
Ind	Tuple	(A_AGE, PEMLR)	Ind	Label (GRP, DIR, PUB)	
0	(0,0)	$(-\infty, 24.01) \times \{2\}$	0,1,2,3,4	NNN, NNY, NY_, YNN, Y1Y	0
1	(1,0)	$(24.01, 55.99) \times \{2\}$	2	NY_	2
2	(2,0)	$(55.99, 64.99) \times \{2\}$	4	Y1Y	1
3	(3,0)	$(64.99, \infty) \times \{2\}$	2	NY_	1
4	(0,1)	$(-\infty, 24.01) \times \{1\}$	0	NNN	7
5	(1,1)	$(24.01, 55.99) \times \{1\}$	3	YNN	23
6	(2,1)	$(55.99, 64.99) \times \{1\}$	3	YNN	4
7	(3,1)	$(64.99, \infty) \times \{1\}$	2	NY_	5
8	(0,2)	$(-\infty, 24.01) \times \{3, 4, 5, 7\}$	1	NNY	6
9	(1,2)	$(24.01, 55.99) \times \{3, 4, 5, 7\}$	0	NNN	9
10	(2,2)	$(55.99, 64.99) \times \{3, 4, 5, 7\}$	2	NY_	7
11	(3,2)	$(64.99, \infty) \times \{3, 4, 5, 7\}$	4	Y1Y	17
12	(0,3)	$(-\infty, 24.01) \times \{0, 6\}$	1	NNY	15
13	(1,3)	$(24.01, 55.99) \times \{0, 6\}$	1	NNY	1
14	(2,3)	$(55.99, 64.99) \times \{0, 6\}$	2	NY_	1
15	(3,3)	$(64.99, \infty) \times \{0, 6\}$	1	NNY	1

Partial Inconsistency (Iteration 9)

Training Instance					Reported		CPLEX		True	
ID	A_AGE	PEMLR	SSYN	Target	Reg	Pred	Reg	Pred	Reg	Pred
8	4	0	0	0	34	1	8	2	10	1
10	12	0	0	0	34	1	8	2	10	1
20	10	0	0	0	34	1	8	2	10	1
21	85	5	1	1	38	4	9	2	11	4
22	74	5	1	1	38	4	9	2	11	4
23	64	5	1	1	38	4	9	2	11	4
24	73	5	1	1	38	4	9	2	11	4
26	5	0	0	1	34	1	8	2	10	1
27	4	0	0	1	34	1	8	2	10	1
28	10	0	0	1	34	1	8	2	10	1
29	54	6	1	1	34	1	8	2	10	1
30	3	0	0	1	34	1	8	2	10	1
33	17	4	1	1	34	1	8	2	10	1
35	77	6	1	1	34	1	8	2	10	1

CPLEX Engine Log

CPLEX Engine Log in Iteration 1 (1/3)

```
<<< setup

Version identifier: 22.1.1.0 | 2022-11-28 | 9160aff4d
CPXPARAM_MIP_Strategy_File                       3
CPXPARAM_MIP_Limits_Solutions                    1
CPXPARAM_TimeLimit                               86400
CPXPARAM_MIP_Limits_TreeMemory                   204800
Tried aggregator 1 time.
MIP Presolve eliminated 402 rows and 800 columns.
MIP Presolve modified 200 coefficients.
Reduced MIP has 4004 rows, 5507 columns, and 22553 nonzeros.
Reduced MIP has 4643 binaries, 11 generals, 0 SOSs, and 0 indicators.
Presolve time = 0.01 sec. (17.75 ticks)
Found incumbent of value -20.000000 after 0.02 sec. (24.01 ticks)
```

CPLEX Engine Log in Iteration 1 (2/3)

```
Root node processing (before b&c):
Real time             =    0.02 sec. (24.25 ticks)
Parallel b&c, 8 threads:
Real time             =    0.00 sec. (0.00 ticks)
Sync time (average)   =    0.00 sec.
Wait time (average)   =    0.00 sec.
------------
Total (root+branch&cut) =    0.02 sec. (24.25 ticks)
```

CPLEX Engine Log in Iteration 1 (3/3)

```
Iteration 1
Bounds on # of cuts = 8 with [3 3 2]
Error = 80 (out of 100 instances)
Accuracy = 20
Solving time = 0.0003894 min (minutes)
Accumulated time = 0.0003894 min (minutes)

Solution status code = 104
LB on error =  -5500
Relative objective gap = 278.999999999

Selected variables:

Number of selected variables = 0 (0 continuous + 0 categorical)
```

CPLEX Engine Log in Iteration 2 (1/5)

```
Version identifier: 22.1.1.0 | 2022-11-28 | 9160aff4d
CPXPARAM_MIP_Strategy_File                       3
CPXPARAM_MIP_Limits_Solutions                    1
CPXPARAM_TimeLimit                               86399.976635986328
CPXPARAM_MIP_Limits_TreeMemory                   204800
Probing time = 0.01 sec. (4.62 ticks)
Cover probing fixed 8 vars, tightened 40 bounds.
Clique table members: 11812.
MIP emphasis: balance optimality and feasibility.
MIP search method: dynamic search.
Parallel mode: deterministic, using up to 8 threads.
Root relaxation solution time = 0.03 sec. (35.79 ticks)
```

CPLEX Engine Log in Iteration 2 (2/5)

```
Nodes                                                 Cuts/
Node  Left     Objective  IInf  Best Integer    Best Bound    ItCnt     Gap

*     0+           0                -20.0000   -5600.0000               ---
0     0        -800.0000   472       -20.0000    -800.0000     1209      ---
0     0        -800.0000   346       -20.0000     Cuts: 512    1987      ---
0     0        -800.0000   651       -20.0000     Cuts: 874    3508      ---
*     0+           0                -28.0000    -800.0000               ---
```

CPLEX Engine Log in Iteration 2 (3/5)

```
GUB cover cuts applied:  29
Clique cuts applied:  10
Cover cuts applied:  51
Implied bound cuts applied:  242
Flow cuts applied:  6
Mixed integer rounding cuts applied:  186
Zero-half cuts applied:  77
Lift and project cuts applied:  7
Gomory fractional cuts applied:  16
```

CPLEX Engine Log in Iteration 2 (4/5)

```
Root node processing (before b&c):
Real time             =    1.78 sec. (1803.05 ticks)
Parallel b&c, 8 threads:
Real time             =    0.00 sec. (0.00 ticks)
Sync time (average)   =    0.00 sec.
Wait time (average)   =    0.00 sec.
------------
Total (root+branch&cut) =    1.78 sec. (1803.05 ticks)
```

CPLEX Engine Log in Iteration 2 (5/5)

```
Iteration 2
Bounds on # of cuts = 8 with [3 3 2]
Error = 72 (out of 100 instances)
Accuracy = 28
Solving time = 0.029740967 min (minutes)
Accumulated time = 0.030130367 min (minutes)

Solution status code = 104
LB on error =  -700
Relative objective gap = 27.571428571

Selected variables:
PEMLR (Categorical)
SS_YN (Categorical)

Number of selected variables = 2 (0 continuous + 2 categorical)
```

Publications

Abstract for Paper 1

Extended Multiclass Support Vector Machines (SVMs) in Nonlinearly Separable Case with Maximum Training Accuracy

This work concerns linear multiclass classification by support vector machines (SVMs) with the one-versus-all (OVA) technique where training data can be nonlinearly separable. The primary and secondary objectives are to minimize misclassification error and to maximize margin sum. The problem of nonconvexity arising from the second task is addressed by minimizing the sum of squared weight norm whose reciprocal is proved to be a reasonable underestimation of margin sum by the Hölder inequality. These goals are achieved by 0-1 mixed integer linear programming (MILP) and 0-1 mixed integer quadratic programming (MIQP). The approximate SVM is also proposed as an alternative to the underestimating SVM to improve time complexity through problem decomposition. The existence of support vectors is theoretically proved. Both methods are compared with soft-margin SVM in terms of training accuracy and execution time.

Abstract for Paper 2

Most Accurate Multiclass Linear Classifiers Producing Disconnected Decision Regions

The work concerns linear multiclass classification where training data can be arbitrary in shape and a decision region is allowed to be disconnected. The main objective is to achieve highest training accuracy, resulting in originally proposed and alternative polygonal classifiers through the detection of an individual misclassified instance and the investigation of majority voting in a decision region respectively. When separating hyperplanes are restricted to cut only one axis, two similar versions of box classifiers which give rectangular decision regions are proposed. Algorithms for finding minimal hyperplanes and cuts while maintaining the same training accuracy are also provided. All four classifiers are developed through 0-1 mixed integer linear programming (MILP). Alternative versions are recommended when training data is very large in number.

Submission History

Paper	Journal	Submit	Date	Status
1	Machine Learning	01 Mar 2023	04 Mar 2023	Reject → Transfer
	Applied Intelligence	13 Mar 2023	02 Apr 2023	Reject → Transfer
	Soft Computing	10 Apr 2023	05 Aug 2023	Reject → Transfer
	SN Computer Science	13 Sep 2023	05 Mar 2025	Under Review
2	Machine Learning	01 Mar 2023	07 Mar 2023	Reject → Transfer
	Annals of Operational Research	14 Mar 2023	20 Feb 2024	Reject → Transfer
	Operational Research	19 Mar 2024	17 Sep 2024	Under Review

Current Statuses

Extended Multiclass Support Vector Machines (SVMs) in Nonlinearly Separable Case with Maximum Training Accuracy

- Journal: SN Computer Science
- Current status: Under review (25 March 2025)
- Preprint DOI: 10.13140/RG.2.2.30648.11526 (ResearchGate)

Most Accurate Multiclass Linear Classifiers Producing Disconnected Decision Regions

- Journal: Operational Research
- Current status: Under review (17 September 2024)
- Preprint DOI: 10.21203/rs.3.rs-2865002/v1 (Research Square)

References (1/2)

Bernstein, D. J., Duif, N., Lange, T., Schwabe, P., and Yang, B.-Y. (2012). High-speed high-security signatures. Journal of cryptographic engineering, 2(2):77–89.

Cebula, R. J. (2006). A further analysis of determinants of health insurance coverage. International Advances in Economic Research, 12(3):382–389.

Dolinsky, A. and Caputo, R. K. (1997). Psychological and demographic characteristics as determinants of women's health insurance coverage. Journal of Consumer Affairs, 31(2):218–237.

Jin, Y., Hou, Z., and Zhang, D. (2016). Determinants of health insurance coverage among people aged 45 and over in china: Who buys public, private and multiple insurance. PLOS ONE, 11(8):1–15.

References (2/2)

Markowitz, M. A., Gold, M., and Rice, T. (1991). Determinants of health insurance status among young adults. Medical care, pages 6–19.

Mulenga, J., Mulenga, M. C., Musonda, K., and Phiri, C. (2021). Examining gender differentials and determinants of private health insurance coverage in zambia. BMC Health Services Research, 21(1):1–11.

Rivest, R. L., Shamir, A., and Adleman, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM, 21(2):120–126.

Ross, B. C. (2014). Mutual information between discrete and continuous data sets. PLOS ONE, 9(2):1–5.