Responsible Machine Learning with Insurance Applications

Christian Lorentzen & Michael Mayer

Autumn 2024

Table of Contents

Organisation

Audience

- bachelor students
- master students
- SAV students

Setting

- 13 lectures
- exercises are integrated in lectures
- oral exam

Prerequisites

- statistics & probability theory
- maximum likelihood theory
- machine learning & supervised learning

Responsible ML

Why responsible? Common risks are:

- ▶ Model does not solve the original goal or task.
- ▶ Missing understanding of the given data.
- ▶ Bias (model & data).
- ▶ Wrong claims about the ability of a model.

"Statistics is about the honest interpretation of data." (Prof. Simon N. Wood)

This lecture aims to provide a toolbox.

- ▶ Model Comparison and Calibration
- ▶ Explainability

Not in this lecture:

- ▶ Data protection law
- ▶ Ethical questions
- ▶ (AI) Fairness
- ▶ MLOps

Applications of ML Models

Actuarial ML models:

- ▶ Pricing models for pure premium and profitability
- ▶ Reserving models for the ultimate claim costs (RBNS and IBNR)
- ▶ Mortality rates / Life tables
- ▶ Fraud detection
- ▶ Conversion and lapse rate
- ▶ . . .

Most methodology works for any kind of supervised model for example:

- ▶ Natural catastrophe (NatCat) models for annual loss
- ▶ Risk models for loss distribution of the company

Decisions are based on actuarial models.

Some Lessons from History

- IBM Watson image recognition (2015): Predicted tag could initially contain unethical associations ("looser").
- Microsoft's Twitter chatbot Tay (2016) gave inflammatory answers like an extremist.
- Google's neural machine translation (2018) gender bias: "doctor" is assigned a *he*, "lazy" is a *she*.
- Apple Card 2019 did provide men higher credit limits than women despite gender not being used by the model.

Measurement

Measurement
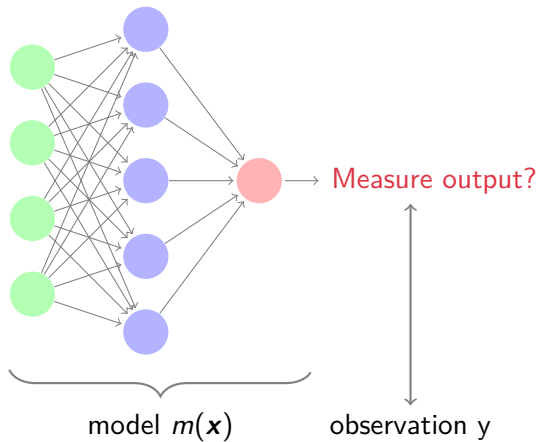

Time

Measurement

Time

Distance

Measurement

Time

Distance

Velocity

model $m(\boldsymbol{x})$ observation y

Measure output?

Exercise 1

▶ What kind of models did you build?

▶ What do you propose to measure to ensure responsible ML usage?

▶ Can you provide examples of problematic or erroneous ML applications?

Table of Contents

Picture of ML

Goal of a model

▶ **inference** on observations/seen data
▶ **prediction** on new, unseen data

Supervised Statistical . . .

Data at population level

▶ Features $\boldsymbol{X}$ take values in some possibly high dimensional feature space $\mathcal{X}$ such as $\mathbb{R}^p$.

▶ Output / response / target variable Y takes values in some space $\mathcal{Y}$, which we assume to be a subset of $\mathbb{R}$ (later called observation domain O).

Remark
We consider both $\boldsymbol{X}$ and Y to be **random variables** with joint probability distribution $F_{\boldsymbol{X},Y}$.

Note
In practice, however, $F_{\boldsymbol{X},Y}$ is usually unknown.

...Learning I

Prediction Goal ❗

- ▶ **Probabilistic** predictions aim for $F_{Y|\boldsymbol{X}}$.
- ▶ **Point** predictions aim for a property / (target) functional $T(F_{Y|\boldsymbol{X}})$.

Remark

Y is random, there is no deterministic function $Y = g(\boldsymbol{X})$.
Convention: $T(F_{Y|\boldsymbol{X}}) = T(Y|\boldsymbol{X})$

Example

- ▶ expectation $= \frac{1}{2}$-expectile, $T(Y|\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$
- ▶ median $= \frac{1}{2}$-quantile
- ▶ α-expectile $T(Y|\boldsymbol{X}) = e_\alpha(Y|\boldsymbol{X})$
  solution e of $\alpha \int_e^\infty (y - e)\, \mathrm{d}F(y) = (1 - \alpha) \int_{-\infty}^e (e - y)\, \mathrm{d}F(y)$
- ▶ α-quantile $T(Y|\boldsymbol{X}) = q_\alpha(Y|\boldsymbol{X})$, any q with $\lim_{y\uparrow q} F(y) \leq \alpha \leq F(q)$
  lower quantile $q_\alpha(Y|\boldsymbol{X}) = \inf\{t \in \mathbb{R} \mid F_{Y|\boldsymbol{X}}(t) \geq \alpha\}$

Data Set

Histogram of UltimateIncurredClaimCost

Workers Compensation data set https://www.openml.org/d/42876
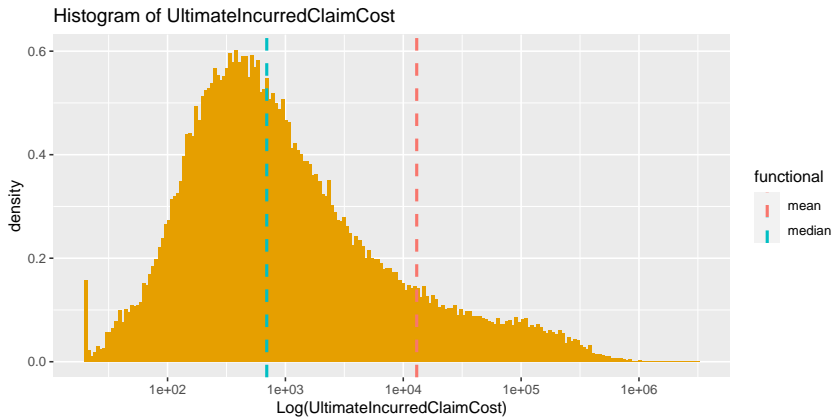
... Learning II

Model

We want to find a model $m \in \mathcal{M}$ from some model class $\mathcal{M}$ to predict $T(Y|\boldsymbol{X})$ by $m(\boldsymbol{X})$.

Loss/Scoring function

▶ We need a **loss** or **scoring function** S to measure the deviation of the model prediction $m(\boldsymbol{X})$ from T using observations Y by $S(m(\boldsymbol{X}), Y)$.

▶ Convention: The smaller S, the better.

▶ For model training as well as model comparison.

Example

▶ squared error
$S(z, y) = (z - y)^2$

▶ absolute error
$S(z, y) = |z - y|$

Iterative optimisation (boosting, gradient descent)

▶ $\overline{S}(m) = \sum_i S(m(\boldsymbol{x}_i), y_i)$

▶ $m_{j+1} \approx \arg\min_{m \in \mathcal{M}} \underbrace{\overline{S}(m) - \overline{S}(m_j)}_{\text{model comparison}}$

Statistical Risk

Statistical risk ❗

The **statistical risk** of model m (under distribution $F_{Y,\boldsymbol{X}}$):

$$R(m) = \mathbb{E}\left[S(m(\boldsymbol{X}), Y)\right] = \mathbb{E}\left[\mathbb{E}\left[S(m(\boldsymbol{X}), Y)|\boldsymbol{X}\right]\right] \tag{1}$$

Ideal model / Bayes rule

$$m^{\star} = \underset{m \in \mathcal{M}}{\arg\min}\, R(m) \tag{2}$$

Note

▶ Use S such that $m^{\star} = T(Y|\boldsymbol{X}) \Rightarrow$ **consistency**

▶ $F_{Y,\boldsymbol{X}}$ and therefore $R(m)$ usually not known.

▶ Neither existence nor uniqueness of $m^{\star}$ are guaranteed.

▶ $m = Y$ is almost surely not $T(Y|\boldsymbol{X})$ and not available at prediction time.

Supervised Learning at Sample Level

Data sample

▶ Observations of i.i.d. input-output pairs $(\mathbf{x}_i, y_i)$ from $F_{Y,\mathbf{X}}$, $i = 1, \ldots, n$
sample of observed y available $\Rightarrow$ term **supervised**

▶ $D = \{(\mathbf{x}_i, y_i), i = 1 \ldots n\}$

Workers Compensation dataset `https://www.openml.org/d/42876`

$y = $ UltimateIncurredClaimCost	InitialCaseEstimate	Age	Gender	WeeklyPay
102	9500	45	M	500
493	1000	18	F	373

Note

▶ Most results are valid without i.i.d. assumption, e.g., forecast comparison explicitly comes from time series.

▶ It is good practise to know your data well $\Rightarrow$ exploratory data analysis (EDA).

Empirical Risk

With the inaccessibility of $R(m)$, one resorts to an estimation by given sample data D.

Empirical risik

$$\overline{R}(m; D) = \overline{S}(m; D) = \frac{1}{n} \sum_{(\mathbf{x}_i, y_i) \in D} S(m(\mathbf{x}_i), y_i) \tag{3}$$

Empirical risk minimisation (ERM) is the actual learning/training step:

$$\widehat{m} = \widehat{m}(\,\cdot\,; D_{\text{train}}) = \underset{m \in \mathcal{M}}{\arg\min}\ \overline{R}(m; D_{\text{train}}) \tag{4}$$

Maximum likelihood estimation

Use a parametric model class $\mathcal{M} = \{m_\theta | \theta \in R^d\}$ and set S to the negative log likelihood function of the assumed data generating process.

Core of the Learning Problem

1. Estimation error
For small training sample size, $\widehat{m}$ has a high (sample) uncertainty.

2. In-Sample risk is biased ❗
The **in-sample risk** or **training loss** $\overline{R}(\widehat{m}(\,\cdot\,; D_{\text{train}}); D_{\text{train}})$, obtained from training and evaluating $\widehat{m}$ on the same data D_{train}, is a biased estimate for $R(\widehat{m})$, usually over-optimistic.

Definition (Overfitting)
Model $m \in \mathcal{M}$ overfits (w.r.t. model complexity given by Ω) the training data D_{train} if there exists another model $m' \in \mathcal{M}$ with $\Omega(m') < \Omega(m)$ such that $\overline{R}(m; D_{\text{train}}) \leq \overline{R}(m'; D_{\text{train}})$, but $R(m) > R(m')$.

Therefore, we always include the trivial model in $\mathcal{M}$ (constant prediction).

Mitigating Overfitting

Adding a penalty

Add a penalty Ω accounting for model complexity/capacity:

$$\underset{m \in \mathcal{M}}{\arg\min} \, \overline{R}(m; D_{\text{train}}) + \lambda \Omega \,. \tag{5}$$

Ridge regression

$\arg\min_{\beta \in \mathbb{R}} \frac{1}{n} \sum_{(\boldsymbol{x}_i, y_i) \in D} (\boldsymbol{x}_i \beta - y_i)^2 + \lambda \|\beta\|_2^2$
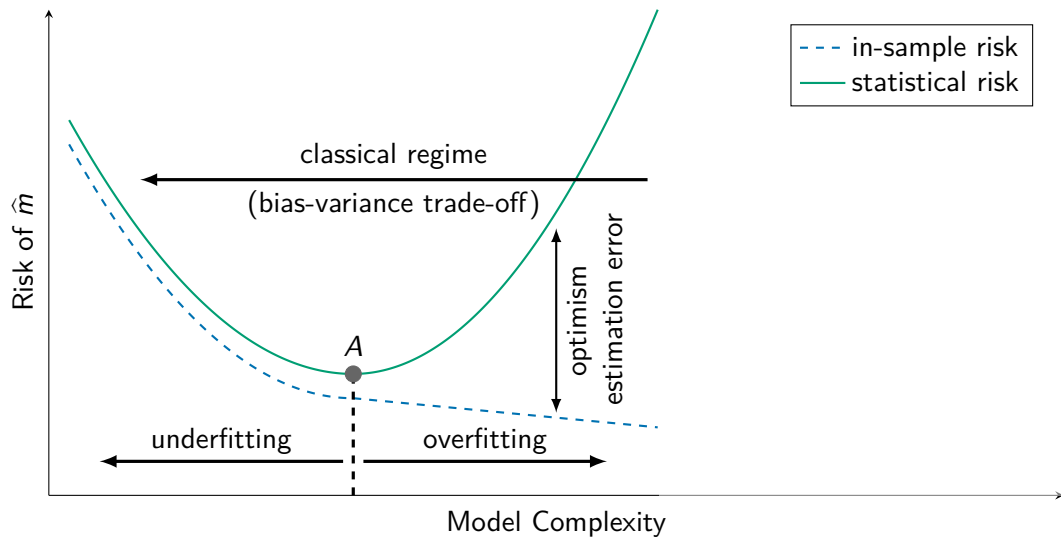
Out-of-sample evaluation

Monitor the **out-of-sample** risk on (ideally) **independent** (and identically distributed) test or validation data D_{test}
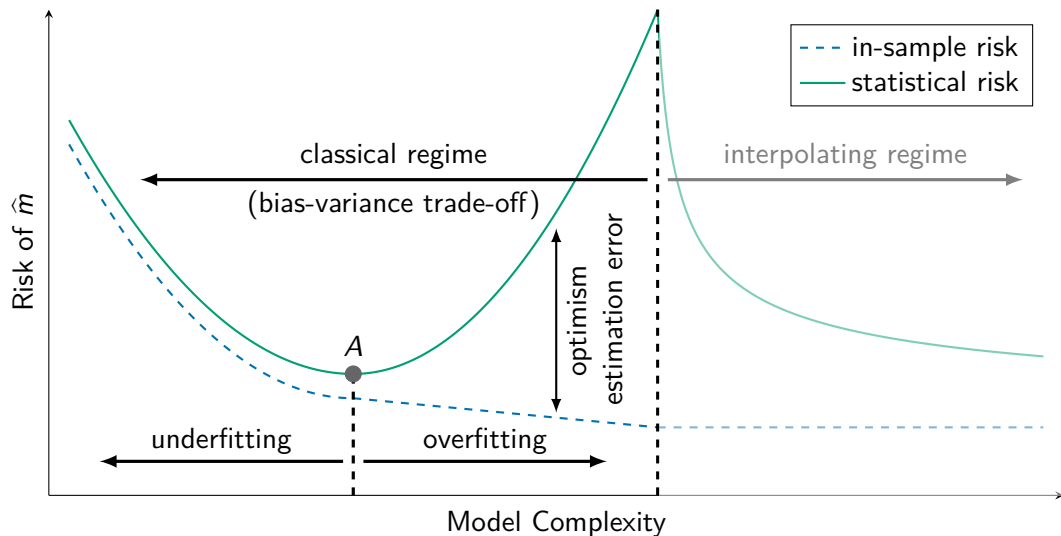
$$\overline{R}(\widehat{m}(\,\cdot\,; D_{\text{train}}); D_{\text{test}}) \,. \tag{6}$$

$\Rightarrow$ need more data!

Statistical Risk vs In-Sample Risk

Statistical Risk vs In-Sample Risk

Decomposition of the Statistical Risk

$$
\begin{aligned}
R(m) = \; & \inf_{g:\mathcal{X}\to\mathcal{Y}} R(g) && \text{inherent unpredictability} && (7) \\
& + \inf_{f\in\mathcal{M}} R(f) && - \inf_{g:\mathcal{X}\to\mathcal{Y}} R(g) && \text{approximation error} \\
& + \inf_{f\in\mathcal{M}} \overline{R}(f;D) && - \inf_{f\in\mathcal{M}} R(f) && \text{estimation error I} \\
& + \overline{R}(m;D) && - \inf_{f\in\mathcal{M}} \overline{R}(f;D) && \text{optimisation error} \\
& + R(m) && - \overline{R}(m;D) && \text{estimation error II}
\end{aligned}
$$

Data Split

Train-Validation-Test-Application Split ⓘ

▶ **Training set** for model fitting, typically the largest set.

▶ **Validation set** for model comparison and model selection. Typically, this set is used to tune a model of a given model class while building (fitting) models on the training set.[1] The result is a "final" model for the given model class that is often refit on the joint training and validation sets.

▶ **Test set** for assessment and comparison of final models. Once the model building phase is finished, this set is used to calculate an unbiased estimate of the statistical risk. It may be used to pick the best one of the (few) final models.

▶ **Application set**. This is the data the model is used for in production. It consists of feature variables only. If the observations of the response become known after a certain time delay, it can serve to monitor the performance of the model.

---

[1]   Examples are variable selection and specification of terms for linear models, finding optimal architecture and early stopping for neural nets, hyperparameter tuning of boosted trees. This way, the validation set is heavily used and therefore does not provide an unbiased performance estimate anymore.

Advice for Data Splits

- **Never ever** look at the test set while still training models.
- The more you use a data set, e.g., the validation set, the less reliable are the results. (analogy: data the new oil $\Rightarrow$ data can be burnt)

Note
A methodological sound train-validation-test split and usage pattern is essential for building good models and for an unbiased assessment of predictive performance.

Cross-Validation
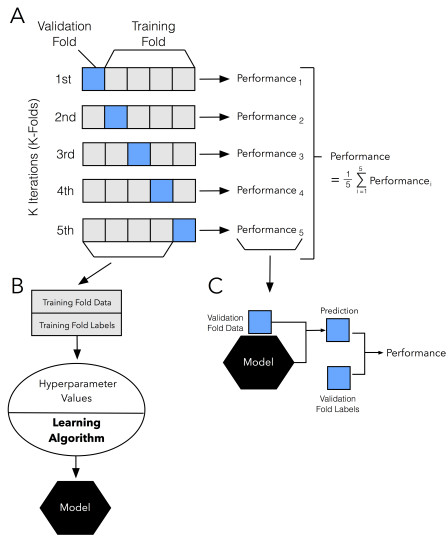
source:
https://sebastianraschka.com/blog/2016/model-evaluation-selection-part3.html

Cross-Validation

CV

- Divide training data into $k = 1, \ldots, K$ disjoint folds D_{train}^k.
- Train model $m_k(\cdot; D_{\text{train}}^{-k})$ on all folds but the kth fold, i.e. on $D_{\text{train}}^{-k} = \bigcup_{j \neq k}^{K} D_{\text{train}}^j$
- Evaluate m_k on fold k as validation set: $\overline{S}_k = \overline{S}(m_k, D_{\text{train}}^k)$.
- Repeat for all k and average the scores: $\overline{S}_{CV} = \frac{1}{K} \sum_k \overline{S}_k$.

Disadvantage

This is an average of the scores of K different models m_k and estimates the expected score over all training sets $\mathbb{E}_{D_{\text{train}}}[R(m(\cdot; D_{\text{train}}))] = \mathbb{E}_{D_{\text{train}}}[\mathbb{E}[S(m(\boldsymbol{X}; D_{\text{train}}), Y)]]$.

Further points

- Many possible splitting schemes
- Account for (dependency) structure in the data
- Possibly different weighting in the final score average

CV Splitting Schemes

Simple splitting schemes

▶ K-fold CV: Divide training data into $K \approx 5 \ldots 10$ equally large folds.

▶ Leave-one-out (LOO): Divide into $K = N$ folds, i.e. only one observation in validation set.
  This enables fast algorithms for linear models (and linear smoothers), but is almost infeasible for other model classes.

▶ Leave-n-out

Ensuring identical distribution

▶ Random shuffling might help to make the folds more identically distributed.

▶ Stratified sampling, in particular for classification problems.
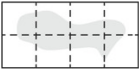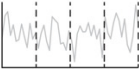 But this might generate dependencies between train and validation set!

CV Data Dependencies
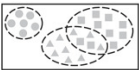
Dependence structure	Parametric solution	Blocking	Blocking illustration
Spatial	Spatial models (e.g.CAR, INLA, GWR)	Spatial	
Temporal	Time-series models (e.g.ARIMA)	Temporal	
Grouping	Mixed effect models (e.g. GLMM)	Group	
Hierarchical / Phylogenetic	Phylogenetic models (e.g. PGLS)	Hierarchical	

Figure: source: doi:10.1111/ecog.02881

Data dependencies

▶ Blocking strategy / Grouped sampling: Same claim or customer ID should only be in one single fold to prevent data leakage from D^{-k} into D^k.

▶ (Spatio-) Temporal structure: Usual assumption is that correlation reduces with (spatio/temporal) distance. For time series: out-of-time or forward-validation scheme:
  ▶ Forecasting horizon: 1-time-step or k-time-steps ahead
  ▶ Fixed vs rolling origin
  ▶ Fixed vs rolling windows

Learning Recipe

Ingredients

- Data sample $D = \{(\mathbf{x}_i, y_i), i = 1 \ldots n\}$
- Chose target functional T.
- Chose a model class $\mathcal{M}$.
- Chose a **loss/scoring function** $S(z_i, y_i) \in \mathbb{R}^p$.

Learn from training data

- Split D into training and validation/test data.
- Train $\hat{m}$ on D_{train}.
- Evaluate on D_{test}

Loss/Scoring function ❗

The loss function should be chosen in line with the directive T in that it should be **strictly consistent** for T.

Exercises I

Exercise 2

Given 2 observations $y = (0, 2)$ and the squared error as loss or scoring function, we furthermore assume constant model prediction $m(\cdot) = z$. Plot the score, e.g., the empirical risk, versus z. What is the optimal constant prediction?

Exercise 3

Proof the following properties of the expectile $e_\alpha(X)$, $X \sim F$ with density f:

▶ $e_\alpha(aX + b) = ae_\alpha(X) + b$ for $a > 0$ and $b \in \mathbb{R}$
▶ $e_\alpha(-X) = -e_{1-\alpha}(X)$

Exercise 4

Given continuous F with finite $\mathbb{E}[Y]$, calculate the Bayes rule for

▶ squared error $S(z, y) = (z - y)^2$, finite 2nd moment of F

Exercises II

- pinball loss $S(z, y) = (\mathbb{1}\{z \geq y\} - \alpha)(z - y)$
- asymmetric piecewise quadratic scoring function (APQSF)
  $S(z, y) = |\mathbb{1}\{z \geq y\} - \alpha|(z - y)^2$ with finite 2nd moment of F.

Exercise 5

Define optimism as $\mathrm{op} = \mathrm{Err}_{in} - \overline{R}_{in}$ with **in-sample risk** $\overline{R}_{in} = \overline{R}(\widehat{m}(\,\cdot\,; D_{\text{train}}); D_{\text{train}})$ and **in-sample error** $\mathrm{Err}_{in} = \frac{1}{n} \sum_{(\boldsymbol{x}_i) \in D_{\text{train}}} \mathbb{E}_{\tilde{Y}_i | X = \boldsymbol{x}_i}[S(m(\boldsymbol{x}_i), \tilde{Y}_i)]$, where $\tilde{Y}_i$ is a new random response value for each $\boldsymbol{x}_i$ in the training set independent of Y_i. Calculate the expected optimism over the responses of the training set, $\omega = \mathbb{E}_{Y_i \in D_{\text{train}}}[\mathrm{op}]$, for the squared error.

Exercise 6

Give examples of possible penalties Ω, for linear, tree based and neural net models.

GLM

Generalised Linear Models estimate the expectation $\mu = \mathbb{E}[Y|\boldsymbol{X}]$.

Building blocks

▶ Numerical features $\boldsymbol{X} \in \mathbb{R}^{n,p}$ and coefficients (or weights) $\beta \in \mathbb{R}^p$
They form the linear predictor $\eta = \boldsymbol{X} \cdot \beta$.

▶ Injective inverse link / response function h: $m(\boldsymbol{X}) = h(\eta)$

▶ Deviance of Exponential Dispersion Family (EDM) as loss function

GLMs are linear in the coefficients β in link space.

Statistical Assumptions

$Y|\boldsymbol{X} \overset{\text{i.i.d}}{\sim} \text{EDM}$ with $f \sim e^{\frac{y\theta - \kappa(\theta)}{\phi}}$, $E[Y] = \kappa'(\theta)$ and $\text{Var}[Y] = \phi\kappa''(\theta)$
For estimation of β, only mean and variance of Y really matter, but ϕ cancels out:

▶ $\mathbb{E}[Y_i|\boldsymbol{X}] = \mu_i$

▶ $\text{Var}[Y_i|\boldsymbol{X}] = \sigma_i^2 = \frac{\phi}{w_i}v(\mu_i)$ with dispersion parameter ϕ, weights w_i and variance function $v(\mu)$ given by the EDM.

Canonical Link Functions

The first order condition for estimation is called **score equation**:

$$0 \overset{!}{=} \boldsymbol{X}'\mathbf{D}\Sigma^{-1}(y - \mu) \tag{8}$$

with $\mathbf{D} = \text{diag}(h'(\eta))$, $\Sigma = \text{diag}(\sigma^2)$.
Canonical link[2] such that $\mathbf{D}\Sigma^{-1}$ cancels out: $h'(\eta) \propto v(h(\eta)) \propto \sigma^2$.
The score equation is then called **balance property** ❗:

$$0 \overset{!}{=} \boldsymbol{X}'(y - \mu) \tag{9}$$

Example

▶ Gauss $v(\mu) = 1 \Rightarrow h'(\eta) = 1 \Rightarrow h(\eta) = \eta$
▶ Poisson $v(\mu) = \mu \Rightarrow h'(\eta) = h(\eta) \Rightarrow h(\eta) = \exp(\eta)$
▶ Tweedie $v(\mu) = \mu^p \Rightarrow h'(\eta) = \frac{1}{1-p}h(\eta)^p \Rightarrow h(\eta) = \eta^{\frac{1}{1-p}}$

---

[2]   Usually defined as $h^{-1}(\mu) = \theta(\mu) = (\kappa')^{-1}(\mu)$.

Neural Net

Architecture

- ▶ Set of connected neurons, often structured in layers.
- ▶ Each neuron
  - ▶ gets as input the outputs $\{z_j, j \in \mathcal{I}\}$ from other neurons input layer gets the features as input $z_j = \boldsymbol{x}_j$
  - ▶ calculates the propagation function $\eta(\boldsymbol{z}) = \sum_{j \in \mathcal{I}} z_j \beta_j + b$ with weights (coefficients) β and bias (intercept) b
  - ▶ outputs $z = \phi(\eta(\boldsymbol{z}))$ with some activation function ϕ (sigmoid/logistic, hyperbolic tangent, ReLU, . . .)
- ▶ Prediction is a nested function $m(\boldsymbol{x}) = z_{output}(\{z_j\})$.

Learning / Training

- ▶ Choose loss/scoring function S.
- ▶ Use optimisation methods for $\arg\min_{\beta, b} \overline{R}(m)$

Input layer   Hidden layer   Output layer
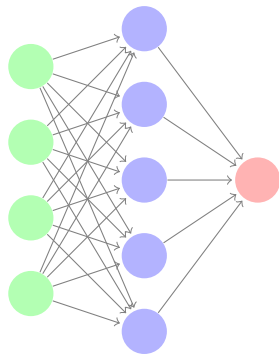
feed forward net

Decision Tree

(Binary) Decision trees are piecewise linear functions. They can model any target T.

Building blocks

▶ Features $\boldsymbol{X}$ with ordering $x_{i_1,j} < x_{i_2,j}$ (observations i_1, i_2, feature j)

▶ Hierarchical structured nodes with feature index j and thresholds θ that partition the feature space $\mathcal{X}$ into disjoint sets Q_k

▶ Loss function S for finding split (j, θ) and node prediction b; plugging in $b = \arg\min_z \overline{S}(z; Q(\theta))$ simplifies loss to entropy, here called splitting criterion, e.g., Gini criterion

▶ K terminal nodes (aka leaves) with predicted value b_k give tree prediction $m(\boldsymbol{x}) = \text{tree}(\boldsymbol{x}, \{b_k, Q_k\}_1^K) = \sum_{k=1}^{K} b_k \mathbb{1}\{\boldsymbol{x} \in Q_k\}$

Decision Trees

root node
$x_2 > 3$

no   yes

$x_1 > 0$ $x_1 > 7$

$x_1 > -5$ $x_3 > -3$

leaf
$b_5 = 2$

leaf
$b_6 = -3$

leaf
$b_1 = 3$

leaf
$b_2 = -2$

leaf
$b_3 = 1$

leaf
$b_4 = 4$
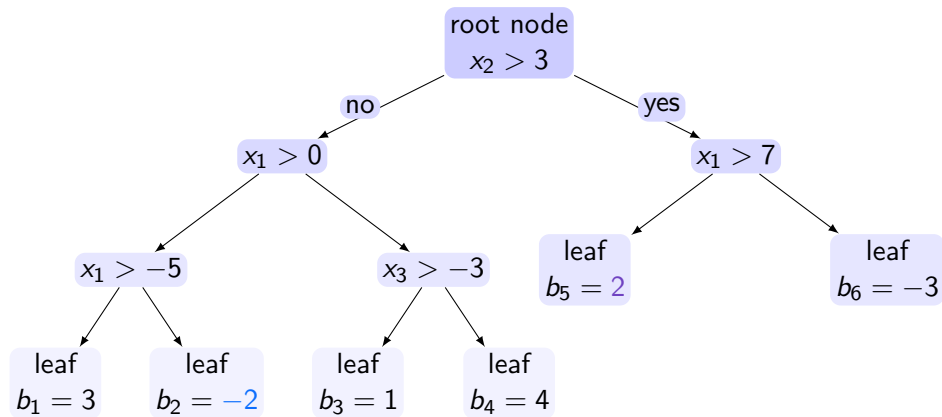
x_1	x_2	x_3	$m(x)$
-4	2	2	-2
-4	4	2	2

Ensemble: Random Forests

Ensemble
Given K models m_k, pool their predictions as $m(\boldsymbol{x}) = \frac{1}{K} \sum_1^K m_k(\boldsymbol{x})$.

Random forest
Fit K different (independent) decision trees, each only on a subset of observations and/or features and pool them together. Each single tree is usually highly overfitted.

Ensemble: Gradient Boosting

Boosting

- Injective inverse link / response function $m(\boldsymbol{x}) = h(F_k(\boldsymbol{x}))$.
- After k fitting stages/iterations we have $F_k(\boldsymbol{x}) = \sum_{j=1}^{k} f_j(\boldsymbol{x})$.
- In each (learning) iteration, add f_k while keeping F_{k-1} fixed.
- Learn f_k by fitting on the residual loss: $\arg\min_{f_k} \overline{R}(h(F_{k-1} + f_k); D_{\text{train}})$.

Gradient boosting (Optimisation in function space)

- Use gradient step: $f_k(\boldsymbol{x}) = -\rho_k g_k(\boldsymbol{x})$
- Gradients $g_k(\boldsymbol{x}) = \left[\frac{\partial \mathbb{E}[S(h(F(\boldsymbol{x})), Y) | \boldsymbol{X}=\boldsymbol{x}]}{\partial F(\boldsymbol{x})} \right]_{F(\boldsymbol{x})=F_{k-1}(\boldsymbol{x})}$ unknown as $F_{Y|\boldsymbol{X}}$ unknown
- But we can evaluate $g_k(\boldsymbol{x})$ on data points[3] $\Rightarrow$ **Fit _stage model_ on response $g_k(\boldsymbol{x})$ via squared error**, call it $\hat{g}_k$, such that $\hat{f}_k = -\rho_k \hat{g}_k \approx -\rho_k g_k$ can be evaluated on all $\boldsymbol{x} \in \mathcal{X}$ (not just training data)
- Line search $\rho_k = \arg\min_\rho R(h(F_{k-1}(\boldsymbol{X}) - \rho \hat{g}_k(\boldsymbol{X})))$

---

[3] The expectation $\mathbb{E}[\cdot | X]$ in g_k is replaced by plugging in the data points (y_i, x_i) instead.

Gradient Boosted Trees

Trees as base learner

- Use trees $\hat{g}_k(\boldsymbol{x}) = \text{tree}(\boldsymbol{x}, \{b_j, Q_j\}_1^J)$.
- On leaf Q_j we have $b_j = -\text{mean}_{i \in Q_j} g_k(\boldsymbol{x}_i)$.
- Line search to find ρ_k.
- Update $F_k(\boldsymbol{x}) = F_{k-1} + \rho_k \sum_{j \in \text{leaves}}^J b_{k,j} \mathbb{1}\{\boldsymbol{x} \in Q_j\}$.

As leaves are disjoint, this can be seen as J separate boosting steps:

$$F_k(\boldsymbol{x}) = F_{k-1} + \sum_{j \in \text{leaves}}^J \tilde{b}_{k,j} \mathbb{1}\{\boldsymbol{x} \in Q_j\}$$

with $\tilde{b}_{k,j} = \rho_k b_{k,j}$. The values $\tilde{b}$ can be found by a line search on each leaf:

$$\tilde{b}_{k,j} = \arg\min_b \sum_{i \in Q_j} S(h(F_{k-1}(\boldsymbol{x}_i) + b), y_i)$$

Note: This gives optimal line search per leaf instead of a "global" line search for the whole tree.

Modern Gradient Boosting

- Hessian (2. order)
    - $h_k(\boldsymbol{x}) = \mathbb{E}\left[\frac{\partial^2 S(h(F(\boldsymbol{x})), Y)}{\partial^2 F(\boldsymbol{x})} | \boldsymbol{X} = \boldsymbol{x}\right]_{F(\boldsymbol{x}) = F_{k-1}(\boldsymbol{x})}$
    - 2. order Taylor of $\overline{R}(F_k; D) \approx const + \frac{1}{2n}\sum_i h_k(\boldsymbol{x}_i)\left(f_k(\boldsymbol{x}_i) + \frac{g_k(\boldsymbol{x}_i)}{h_k(\boldsymbol{x}_i)}\right)^2$
    - Optionally add penalty $\Omega = \frac{\lambda}{2}\sum_{\text{leaves } j} b_j^2$.
    - Fit a tree via weighted least squares on response $-g/h$ with weights h.
    - On leaf Q_j, we have constant prediction $b_j = -\frac{\sum_{i \in Q_j} g_k(\boldsymbol{x}_i)}{\sum_{i \in Q_j} h_k(\boldsymbol{x}_i) + \lambda}$
- Histogram: Calculate histogram for each feature and accumulate h and g/h.
- Categorical (nominal) features: Reduces splitting from $\mathcal{O}(2^K)$ to $\mathcal{O}(K\log(K) + K)$ by Fisher's algorithm[4]
    - Sort by g/h
    - Treat them as continuous/ordinal

---

[4]   W.D. Fisher (1958) "On Grouping for Maximum Homogeneity" Journal of the American Statistical Association, 53, 789-798.

Exercises I

Exercise 7

Given linear models with response function h and prediction $m(\boldsymbol{x}) = h(\boldsymbol{x} \cdot \beta)$, derive the first order optimality condition with Bregman functions

$$S(z, y) = \phi(y) - \phi(z) + \phi'(z)(z - y), \quad \text{convex } \phi$$

as loss when training for β. Which condition has to hold to arrive at the balance property (9)? The resulting link function is the **canonical** one.

Exercise 8

Derive the splitting criterion of a tree node for loss/scoring functions:

1. squared error
2. log loss $S(z, y) = -y \log(z) - (1 - y) \log(1 - z)$, $y \in \{0, 1\}$, $z \in [0, 1]$

Exercises II

Exercise 9

For median regressing gradient boosted trees using the absolute error, derive:

1. the "global" line search ρ_k
2. the per leave line search $\tilde{b}_{k,j}$

Exercise 10

Fit a GLM, a random forest and a gradient boosted tree model on the Workers Compensation dataset https://www.openml.org/d/42876. Model goal is $\mathbb{E}[\text{UltimateIncurredClaimCost}|\boldsymbol{X}]$.

Table of Contents

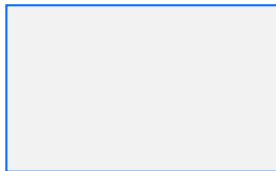# Measuring Predictive Model Performance

Time

Measuring Predictive Model Performance

 Time

$m_1(\boldsymbol{x})$ better than $m_2(\boldsymbol{x})$?

Measuring Predictive Model Performance

Time

Stricly Consistent Scoring Function S

$m_1(\boldsymbol{x})$ better than $m_2(\boldsymbol{x})$?

Notation

- **Observation domain** O, which comprises the potential outcomes of a future observation.
- Convex class $\mathcal{F}$ of probability measures on the observation domain O (equipped with a suitable σ-algebra), which constitutes a family of probability distributions for the future observation.
- **Action domain** A, which comprises the potential actions of a decision maker.
- **Scoring or loss function** $S : A \times O \to \mathbb{R}$, where $S(a, o)$ represents the monetary or societal cost when the decision maker takes the action (or point forecast) $a \in A$ and the observation $o \in O$ materialises.
- A **scoring rule** is a function $\mathbf{S} : \mathcal{F} \times O \to \mathbb{R}$, $\mathbf{S}(F, o)$ is the penalty for probabilistic prediction $F \in \mathcal{F}$ and observation o.
- Statistical functional $T : \mathcal{F} \to D$, potentially set-valued.

Assumption

Common domain $D = O = A \subseteq \mathbb{R}^d$,

Note: We could have chosen $S : A \times O \to [0, \infty)$, w.l.o.g.

Scoring Functions

Repetition

▶ A scoring function S measures the deviation of the model prediction $m(\boldsymbol{X})$ from T using observations Y: $S(m(\boldsymbol{X}), Y)$.

▶ Convention: The smaller S, the better.

Purpose

▶ Assess predictive performance of the predictions of a model.

▶ Compare the predictiveness of different models.

Scoring rules

A scoring rule **S** is in principle the same as a scoring function but for probabilistic predictions: model goal is F (or pdf f).

Model Comparison

We estimate the expected score $\mathbb{E}[S(m(\boldsymbol{X}), Y)]$, earlier called statistical risk $R(m)$, as

$$\overline{S}(m; D) = \frac{1}{n} \sum_{(\boldsymbol{x}_i, y_i) \in D} S(m(\boldsymbol{x}_i), y_i), \tag{10}$$

which we called empirical risk before. Model m_A is deemed to have an inferior predictive performance than model m_B in terms of the score S (and on the sample D) if

$$\overline{S}(m_A; D) - \overline{S}(m_B; D) = \frac{1}{n} \sum_{(\boldsymbol{x}_i, y_i) \in D} S(m_A(\boldsymbol{x}_i), y_i) - S(m_B(\boldsymbol{x}_i), y_i) > 0. \tag{11}$$

Statistical test

▶ With i.i.d. data, simple t-test.
▶ With (time / serial) correlation, Diebold-Mariano test.

Consistency & Elicitability

Definition (Consistency ❗)

Let $\mathcal{F}$ be a class of probability distributions where the functional T is defined on. A scoring function $S(z, y)$ is a function in a forecast z and an observation y. It is $\mathcal{F}$-**consistent** for T if

$$\int S(t, y)\, \mathrm{d}F(y) \leq \int S(z, y)\, \mathrm{d}F(y) \qquad \text{for all } t \in T(F),\ z \in \mathrm{D},\ F \in \mathcal{F}. \qquad (12)$$

The score is **strictly** $\mathcal{F}$-consistent for T if it is $\mathcal{F}$-consistent for T and if equality in (12) implies that $z \in T(F)$.

Definition (Elicitability ❗)

A functional T is **elicitable** on $\mathcal{F}$ if there is a strictly $\mathcal{F}$-consistent scoring function for it.

Why Consistency Matters?

Consistency

- ▶ It ensures that we get what we want: $m^\star = T(Y|\boldsymbol{X})$.
- ▶ At least in the large sample limit (Law of Large Numbers arguments).
- ▶ Compare with a repeated game where each forecaster gets penalty / loss $S(z, y)$.

Counter example: Use of absolute error $|z - y|$ when we aim for the expectation.

Elicitability

- ▶ Tells us if there exists a consistent scoring function for the functional T.
- ▶ Model comparison and backtesting is (partially) pointless for non-elicitable T.

Counter examples: Mode (for general F), variance (alone) and expected shortfall (alone) are not elicitable.

Proper Scoring Rules

Definition (Propriety)

The scoring rule **S** is **proper** relative to the class $\mathcal{F}$ if

$$\mathbb{E}_G[\mathbf{S}(G, Y)] \leq \mathbb{E}_G[\mathbf{S}(F, Y)]$$

for all $F, G \in \mathcal{F}$. It is **strictly proper**, if equality holds (if and) only if $F = G$.

Theorem (Gneiting 2011 Theorem 3)

Suppose that the scoring function S is $\mathcal{F}-$consistent for the functional T. For each $F \in \mathcal{F}$, let $t_F \in T(F)$. Then $\mathbf{S}(F, y) = S(t_F, y)$ is a proper scoring rule relative to $\mathcal{F}$.

Examples of proper scoring rules

- quadratic score $\mathbf{S}(F, y) = -2f(y) + \int f^2(x)\,\mathrm{d}x$
- logarithmic score $\mathbf{S}(F, y) = -\log f(y) \Rightarrow$ compare MLE
- continuous ranked probability score (CRPS)
  $\mathbf{S}(F, y) = \int (F(x) - \mathbb{1}\{y \leq x\})^2\,\mathrm{d}x = \mathbb{E}_F[|Y - y|] - \frac{1}{2}\,\mathbb{E}_F[|Y - Y'|],\ Y, Y' \overset{\text{i.i.d}}{\sim} F$
- Dawid-Sebastiani score $\mathbf{S}(F, y) = \frac{(y - \mu_F)^2}{\sigma_F^2} + 2\log \sigma_F$ with $\mu_F = \mathbb{E}_F[X]$, $\sigma_F^2 = \mathsf{Var}_F[X]$

Order Sensitivity

Given a one-dimensional, real-valued T ($D \subseteq \mathbb{R}$).

Definition
T is $\mathcal{F}$-**order sensitive** if for any $F \in \mathcal{F}$ and any $z_1, z_2 \in A$ with either $z_1 > z_2 > T(F)$ or $z_1 < z_2 < T(F)$ one has $\mathbb{E}_F[S(z_1, Y)] > \mathbb{E}_F[S(z_2, Y)]$.

Implications
- Order sensitivity of S implies consistency.
- (Under weak regularity conditions:) Strict consistency of S implies order sensitivity.

Note: For T with $D \in \mathbb{R}^k$, different notions of order sensitivity arise, e.g., component-wise order sensitivity.

Convex Level Sets

Theorem (Osband 1985)

If a one-dimensional ($D \subseteq \mathbb{R}$) functional T is elicitable, then its level sets are convex in the following sense: If $F_0, F_1 \in \mathcal{F}$ and $p \in (0,1)$ are such that $F_p = pF_0 + (1-p)F_1 \in \mathcal{F}$, then $t \in T(F_0)$ and $t \in T(F_1)$ imply $t \in T(F_p)$.

Proof.

For $t \in T(F_0)$ and $t \in T(F_1)$, we have $\mathbb{E}_F[S(t,Y)] \leq \mathbb{E}_F[S(z,Y)]$ for all $z \in A$ and $F \in \{F_0, F_1\}$. Then $\mathbb{E}_{F_p}[S(t,Y)] = p\,\mathbb{E}_{F_0}[S(t,Y)] + (1-p)\,\mathbb{E}_{F_1}[S(t,Y)] \leq p\,\mathbb{E}_{F_0}[S(z,Y)] + (1-p)\,\mathbb{E}_{F_1}[S(z,Y)] = \mathbb{E}_{F_p}[S(z,Y)].$ □

Application

Proof that variance is not elicitable:

We have $\text{Var}[\delta_x] = \text{Var}[\delta_y] = \text{Var}_{\delta_y}[Y] = 0$. But

$\text{Var}[p\delta_x + (1-p)\delta_y] = \mathbb{E}_{p\delta_x + (1-p)\delta_y}[(Y - (px + (1-p)y))^2] = p((1-p)x - (1-p)y)^2 + (1-p)(px - py)^2 = p(1-p)(x-y)^2 \neq 0$

Note: The mode has convex level sets, but fails to be elicitable for the class $\mathcal{F}$ of strictly unimodal distributions with continuous Lebesgue density.

Revelation Principle

Theorem (Osband 1985)

Suppose that the class $\mathcal{F}$ is concentrated on the domain D, and let $g : \mathrm{D} \to \mathrm{D}$ be a one-to-one mapping. Then the following holds.

1. *If T is elicitable, then $T_g = g \circ T$ is elicitable.*
2. *If S is consistent for T, then the scoring function $S_g(x, y) = S(g^{-1}(x), y)$ is consistent for T_g.*
3. *If S is strictly consistent for T, then S_g is strictly consistent for T_g.*

Scoring Functions with Weighted Densities

Some Assumptions

- ▶ Functional T is defined on class $\mathcal{F}$ of probability distributions which admit a density, f, with respect to some dominating measure on the domain D.
- ▶ Weight function $w : D \to [0, \infty)$
- ▶ $\mathcal{F}^{(w)} \subseteq \mathcal{F}$ denotes subclass of probability distributions in $\mathcal{F}$ which are such that $\int_D w(y)f(y)\,dy < \infty$, and the probability measure $F^{(w)}$ with density proportional to $w(y)f(y)$ belongs to $\mathcal{F}$. On this subclass $\mathcal{F}^{(w)}$, define the functional

$$T^{(w)} : \mathcal{F}^{(w)} \to I \subseteq \mathbb{R} \qquad F \to T^{(w)}(F) = T(F^{(w)}).$$

Theorem (Gneiting 2011 Theorem 5)

Given the above assumptions, the following holds.

1. *If T is elicitable, then $T^{(w)}$ is elicitable.*
2. *If S is consistent for T relative to $\mathcal{F}$, then $S^{(w)}(z, y) = w(y)S(z, y)$ is consistent for $T^{(w)}$ relative to $\mathcal{F}^{(w)}$.*

Characterisation

Expectation: Bregman functions ❗

$$S(z, y) = \phi(y) - \phi(z) + \phi'(z)(z - y) + a(y) \tag{13}$$

with (strictly) convex ϕ and arbitrary a are (strictly) consistent for $T = \mathbb{E}$

Quantiles: generalised piecewise linear (GPL)

$$S(z, y) = (\mathbb{1}\{y \leq z\} - \alpha)(g(z) - g(y)) + a(y) \tag{14}$$

with (strictly) increasing g and arbitrary a are (strictly) consistent for $T = q_\alpha$

Expectiles

$$S(z, y) = 2|\mathbb{1}\{y \leq z\} - \alpha|(\phi(y) - \phi(z) + \phi'(z)(z - y)) + a(y) \tag{15}$$

with (strictly) convex ϕ and arbitrary a are (strictly) consistent for $T = e_\alpha$.

Mode: zero-one loss

$$S(z, y) = \lambda \mathbb{1}\{z \neq y\} + a(y) \quad \lambda > 0 \tag{16}$$

is strictly consistent for categorical $Y \in \{0, \ldots, k - 1\}$.

Examples of Strictly Consistent Scoring Functions

Functional	Scoring Function	Formula $S(z,y)$	Domain
expectation	squared error	$(y-z)^2$	$y, z \in \mathbb{R}$
	Poisson deviance	$2\left(y \log \frac{y}{z} + z - y\right)$	$y \geq 0, z > 0$
	Gamma deviance	$2\left(\log \frac{z}{y} + \frac{y}{z} - 1\right)$	$y, z > 0$
	Tweedie deviance	$2\left(\frac{y^{2-p}}{(1-p)\cdot(2-p)}\right.$	$y, z > 0$
	$\quad p \in \mathbb{R} \setminus \{1,2\}$	$\left. -\frac{y\cdot z^{1-p}}{1-p} + \frac{z^{2-p}}{2-p}\right)$	$y \geq 0$ for $p < 2$
	homogeneous score	$\lvert y \rvert^a - \lvert z \rvert^a$	$y, z \in \mathbb{R}$
	$\quad a > 1$	$-a\,\mathrm{sign}(z)\lvert z \rvert^{a-1}(y-z)$	
	log loss	$-y \log z - (1-y)\log(1-z)$	$0 \leq y \leq 1$
		$+ y \log y + (1-y)\log(1-y)$	$0 < z < 1$
α-expectile	APQSF[5]	$\lvert \mathbb{1}\{z \geq y\} - \alpha \rvert (z-y)^2$	$\mathbb{R}$
median	absolute error	$\lvert y - z \rvert$	$\mathbb{R}$
α-quantile	pinball loss	$(\mathbb{1}\{z \geq y\} - \alpha)(z-y)$	$\mathbb{R}$

---

[5] asymptotic piecewise quadratic scoring function

Scoring Functions with Weighted Densities

Example

▶ On D $= (0, \infty)$, $S(z, y) = |z^{-\beta} - y^{-\beta}|$ and $w(y) = y^{\beta}$ produce

$$S_{\beta}(z, y) = \left| 1 - \left(\frac{y}{z} \right)^{\beta} \right| \tag{17}$$

▶ $S(z, y)$ is consistent for the median, see Eq. (14) with $g(x) = \text{sign}(b)x^b$.
▶ By Theorem (Gneiting 2010 Th. 5), $S_{\beta}(z, y)$ is consistent for the β-median, $\text{med}^{(\beta)}(F)$, i.e. the median of the distribution with density proportional to $y^{\beta}f(y)$, and f the density of F.

Special cases

▶ $\beta = -1$: absolute percentage error (APE) $S_{-1}(z, y) = |\frac{z-y}{y}|$
▶ $\beta = 1$: relative error (RE) $S(z, y) = |\frac{z-y}{z}|$

Elementary Scoring Functions

With identification function V for quantile or expectile T, see (22) on slide 71, the elementary scoring function

$$S_\theta(z, y) = (\mathbb{1}\{\theta \leq z\} - \mathbb{1}\{\theta \leq y\}) \, V(\theta, y) \tag{18}$$

is consistent for T.

Any (strictly) consistent scoring function admits a mixture representation

$$S(z, y) = \int S_\theta(z, y) \, \mathrm{d}H(\theta) + a(y) \tag{19}$$

for non-negative (positive[6]) measure H on $\mathbb{R}$, with $\mathrm{d}H(\theta) = \mathrm{d}g(\theta)$ for quantiles and $\mathrm{d}H(\theta) = \mathrm{d}\phi'(\theta)$ for expectiles.

Note: $V(z, y) = z - y$ for $T = \mathbb{E}$.

---

[6] H gives positive measure to every non-degenerate interval.
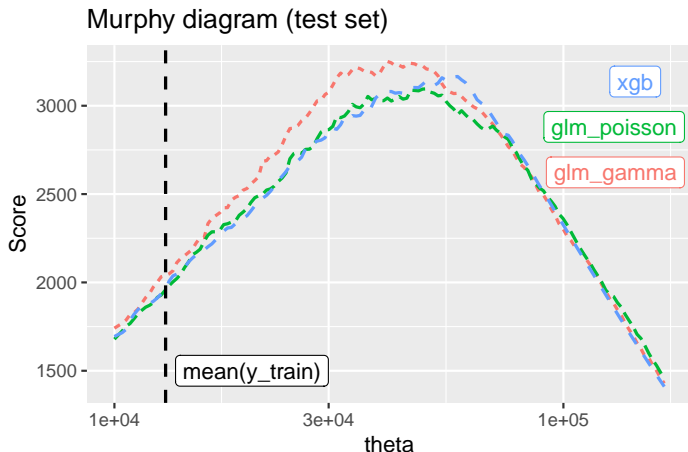
Forecast Dominance

Definition

Prediction/forecast z_1 dominates z_2 if $\mathbb{E}[S(z_1, Y)] < \mathbb{E}[S(z_2, Y)]$ for all (strictly) consistent scoring functions.

Quantiles and expectiles

For quantiles and expectiles this is equivalent to $\mathbb{E}[S_\theta(z_1, Y)] < \mathbb{E}[S_\theta(z_2, Y)]$ for all $\theta \in \mathbb{R}$.

Murphy Diagram

Compare many scoring functions (sliding parameter θ) at once.
Assess forecast dominance.

Murphy diagram (test set)

Elementary scoring function for $\mathbb{E}$: $S_\theta(z, y) = |\theta - y| \mathbb{1}\{\min(z, y) \le \theta < \max(z, y)\}$

Which One to Choose?

Use a strictly consistent scoring function!

But: Which one out of the infinitely many ones (for elicitable T)?

Further criteria

- ▶ Domain / Range of target Y.
- ▶ Degree of homogeneity: $S(tz, ty) = t^h S(z, y)$ for all $t > 0$ and for all z, y
- ▶ Efficiency: How fast is the large sample convergence?
- ▶ Forecast dominance: Is one model dominating for many/all scoring functions? Assess with Murphy diagrams.
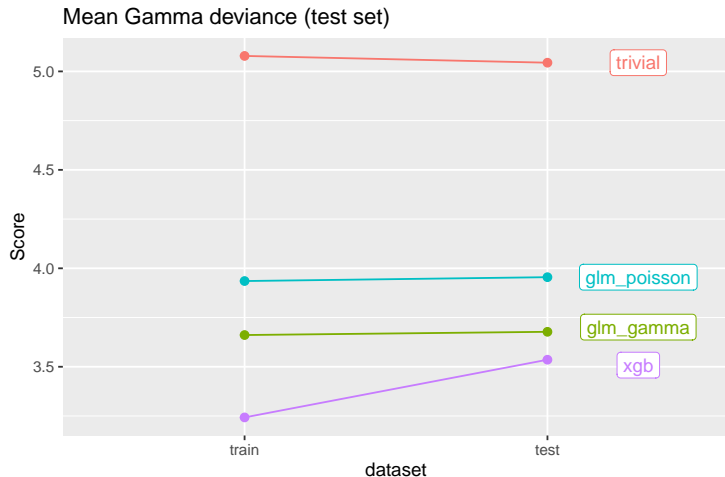
Squared error: $h = 2$

Gamma deviance:
Degree of homogeneity is $h = 0 \Rightarrow$ It only cares about relative differences:
$S(1, 10) = S(10, 100) = S(100, 1000) = 13.39$

Model Comparison

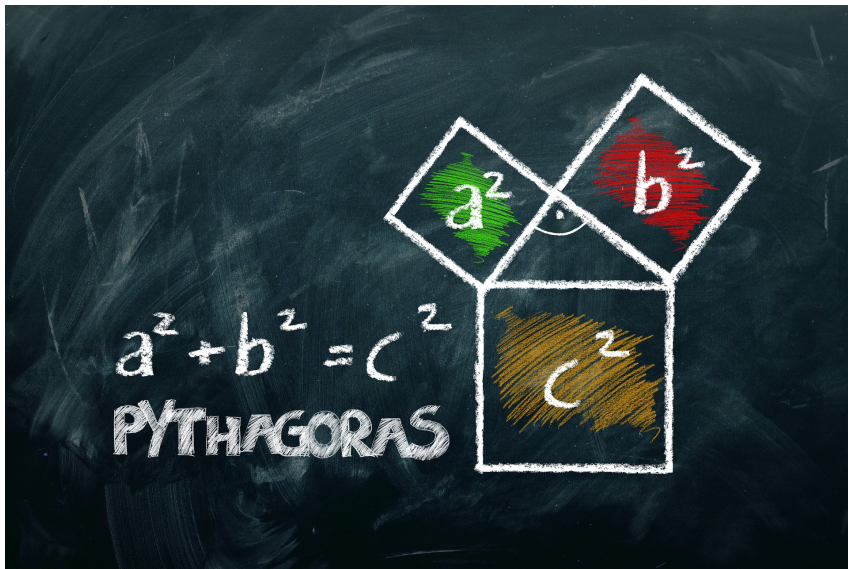Compare empirical mean scores: $\overline{S}(m) = \frac{1}{n} \sum_i S(m(\boldsymbol{x}_i), y_i)$

Gamma deviance for workers compensation

Mean Gamma deviance (test set)

Models:

1. Trivial model always predicts mean(y) of the training set.

2. Poisson GLM with canonical log-link.

3. Gamma GLM with log-link.

4. XGBoost model with Gamma deviance and log-link.

Additive Score Decomposition

Score Decomposition

$$\mathbb{E}[S(m(\boldsymbol{X}), Y)] = \Big\{ \underbrace{\mathbb{E}[S(m(\boldsymbol{X}), Y)] - \mathbb{E}[S(T(Y|\boldsymbol{X}), Y)]}_{\text{conditional miscalibration}} \Big\} \tag{20}$$

$$- \Big\{ \underbrace{\mathbb{E}[S(T(Y), Y)] - \mathbb{E}[S(T(Y|\boldsymbol{X}), Y)]}_{\text{conditional resolution / conditional discrimination}} \Big\} + \underbrace{\mathbb{E}[S(T(Y), Y)]}_{\text{uncertainty / entropy}}$$

can be estimated
$$\begin{cases} = \Big\{ \underbrace{\mathbb{E}[S(m(\boldsymbol{X}), Y)] - \mathbb{E}[S(T(Y|m(\boldsymbol{X})), Y)]}_{\text{auto-miscalibration}} \Big\} \\ - \Big\{ \underbrace{\mathbb{E}[S(T(Y), Y)] - \mathbb{E}[S(T(Y|m(\boldsymbol{X})), Y)]}_{\text{auto-resolution / auto-discrimination}} \Big\} + \underbrace{\mathbb{E}[S(T(Y), Y)]}_{\text{uncertainty / entropy}} \end{cases}$$

Note: Minimising consistent scores amounts to **jointly** minimising miscalibration and maximising resolution!

Squared Error / Brier Score

$$\mathbb{E}[(m(\boldsymbol{X}) - Y)^2] = \underbrace{\mathbb{E}[(m(\boldsymbol{X}) - \mathbb{E}[Y|\boldsymbol{X}])^2]}_{\text{conditional miscalibration}} - \underbrace{\text{Var}[\mathbb{E}[Y|\boldsymbol{X}]]}_{\text{conditional resolution}} + \underbrace{\text{Var}[Y]}_{\text{uncertainty}} \tag{21}$$

Score Decomposition of Gamma Deviance

Again for workers compensation

Model	Mean deviance	Auto-miscalibration	Auto-resolution	Uncertainty
Trivial	5.04	0	0	5.04
GLM Gamma	3.68	0.190	1.56	5.04
GLM Poisson	3.95	0.482	1.57	5.04
XGB	**3.54**	**0.124**	**1.63**	5.04

Isotonic regression

For $T = \mathbb{E}$, one can estimate $\mathbb{E}[Y|m(\boldsymbol{x})]$ by isotonic regression (PAV algorithm) of y_i against $m(\boldsymbol{x}_i)$.

New results[7] show how to extend PAV to quantiles, expectiles and more.

[7] A.I. Jordan, A. Mühlemann & J.F. Ziegel (2022) "Characterizing the optimal solutions to the isotonic regression problem for identifiable functionals" Ann Inst Stat Math 74, 489–514. doi:10.1007/s10463-021-00808-0

Exercises

Exercise 11

Compute the Bayes rule for the scoring functions in Eq. (13), (14) and (15).
Remember Ex. 4. Hint for (14) (difficult): Case distinction $z > q_\alpha$ and $z < q_\alpha$.

Exercise 12

Device a betting game with a wager $\rho_L > 0$ and pay-off scheme depending on a
random outcome Y such that the optimal strategy in expectation is a quantile. Hint:
Have a look at the elementary scoring function.

Exercise 13

Derive the decomposition of the squared error in Eq. (21).

Exercise 14

Calculate the score decomposition of the Gamma deviance for your models on the
Workers Compensation dataset.

Table of Contents

Figure: Scale calibration
source: https://ssec-epos.co.uk/hardware/weighing-scale-calibration-and-service/

Motivation for Calibration

▶ Is the model fit for its prediction task?

▶ How well does the predictions align with observations?

▶ Detect bias and discrimination.

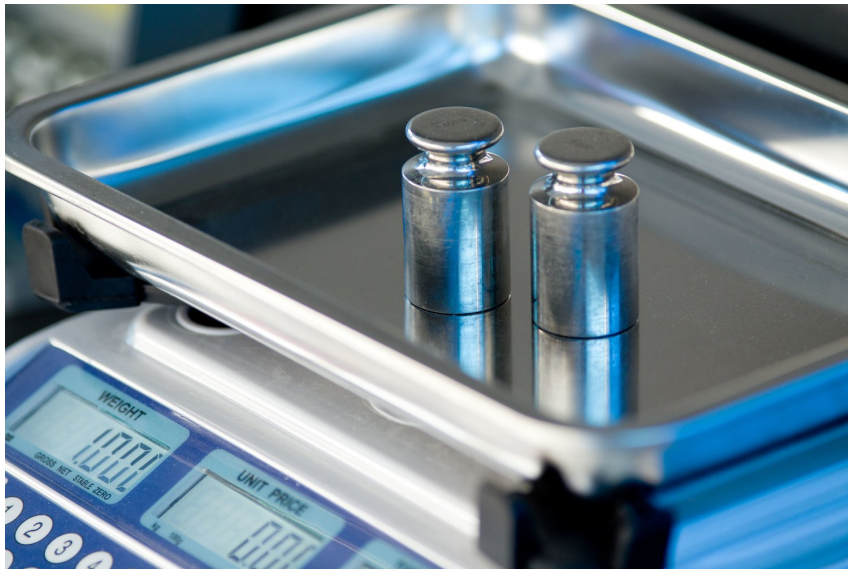Bias can result in bad news.

Motivation for Calibration

- ▶ Is the model fit for its prediction task?
- ▶ How well does the predictions align with observations?
- ▶ Detect bias and discrimination.

Bias can result in bad news.

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Figure: ProPublica article on COMPAS.

Measuring Model Calibration
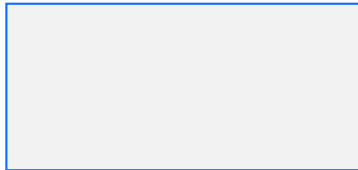
Distance

Is $m(\boldsymbol{x})$ calibrated?

Measuring Model Calibration

Distance

Strict Identification Function V

Is $m(\boldsymbol{x})$ calibrated?

2 Notions of Calibration

Definition

Given a feature–response pair $(\boldsymbol{X}, Y)$, the model $m(\boldsymbol{X})$ is **conditionally calibrated** for the functional T if

$$m(\boldsymbol{X}) = T(Y|\boldsymbol{X}) \qquad \text{almost surely.}$$

The model $m(\boldsymbol{X})$ is **auto-calibrated** for T if

$$m(\boldsymbol{X}) = T(Y|m(\boldsymbol{X})) \qquad \text{almost surely.}$$

Note

▶ Calibration uncovers the use of information.

▶ Conditional calibrated models use information ideally and predict **oracle function** $\boldsymbol{x} \rightarrow T(Y|\boldsymbol{X} = \boldsymbol{x})$.

▶ The trivial model $m(\boldsymbol{X}) = T(Y)$ is auto-calibrated, but a.s. not conditionally calibrated.

▶ **Unconditional calibration** does not have such a definition in terms of T and $m(x)$.

Identification Functions

Definition (❶)

Let $\mathcal{F}$ be a class of probability distributions where the functional T is defined on. A **strict $\mathcal{F}$-identification function** for T is a function $V(z, y)$ in a forecast z and an observation y such that

$$\int V(z, y)\, dF(y) = 0 \quad \Longleftrightarrow \quad z \in T(F) \qquad \text{for all } z \in \mathbb{R},\ F \in \mathcal{F}. \qquad (22)$$

If only the implication $\Longleftarrow$ in (22) holds, then V is just called an $\mathcal{F}$-identification function for T. If T admits a strict $\mathcal{F}$-identification function, it is **identifiable** on $\mathcal{F}$.

Canonical strict identification functions

Functional	Strict Identification Function	Domain of y, z
expectation $\mathbb{E}[Y]$	$V(z, y) = z - y$	$\mathbb{R}$
α-expectile	$V(z, y) = 2|\mathbb{1}\{z \geq y\} - \alpha|(z - y)$	$\mathbb{R}$
median $F_Y^{-1}(0.5)$	$V(z, y) = \mathbb{1}\{z \geq y\} - {}^1/_2$	$\mathbb{R}$
α-quantile $F_Y^{-1}(\alpha)$	$V(z, y) = \mathbb{1}\{z \geq y\} - \alpha$	$\mathbb{R}$

Osbands's Principle for Scoring Functions

Assume $\mathbb{P}(Y = a) = p$ and $\mathbb{P}(Y = b) = 1 - p$. Then, for any S consistent for T and smooth in its first argument, the expected score $\epsilon(c) = pS(c, a) + (1 - p)S(c, b)$ is minimised at $c = t$ fulfilling

$$\epsilon'(t) = pS_{(1)}(t, a) + (1 - p)S_{(1)}(t, b) = 0 \, .$$

Furthermore, (22) implies

$$pV(t, a) + (1 - p)V(t, b) = 0 \, .$$

Combining them gives for all pairwise distinct a, b and $t \in D$

$$S_{(1)}(t, a)/V(t, a) = S_{(1)}(t, b)/V(t, b) = \text{function in } t \text{ alone}$$

and we can write, for some function $h : D \to D$

$$S_{(1)}(z, y) = h(z)V(z, y) \, . \tag{23}$$

Osbands's Principle for Identification Functions

Characterisation[8]
Subject to mild regularity conditions on $\mathcal{F}$ and V:
If V is a strict identification function for $T : \mathcal{F} \to A \subseteq \mathbb{R}^k$, then

$$\{h(z)V(z,y)|h : A \to \mathbb{R}^{k,k}, \det h(z) \neq 0 \text{ for all } z \in A\} \qquad (24)$$

is the entire class of strict identification functions for T.

Identifiability and elicitability
Under some richness assumptions on the class $\mathcal{F}$ and continuity assumptions on T:
For one-dimensional functionals T, identifiability and elicitability are equivalent.

---

[8] T. Dimitriadis, T. Fissler and J. Ziegel. "Osband's Principle for Identification Functions". arxiv:2208.07685

Identification Functions and Calibration

Let V be any strict $\mathcal{F}$-identification function for T.

Conditional calibration

Suppose that $\mathcal{F}$ contains the conditional distributions $F_{Y|\boldsymbol{X}=\boldsymbol{x}}$ for almost all $\boldsymbol{x} \in \mathcal{X}$. Application of (22) to these conditional distributions yields that $m(\boldsymbol{x}) \in T(Y|\boldsymbol{X} = \boldsymbol{x})$ if and only if $\int V(m(\boldsymbol{x}), y)\, \mathrm{d}F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = 0$. This shows that m is conditionally calibrated for T if and only if

$$\mathbb{E}[V(m(\boldsymbol{X}), Y)|\boldsymbol{X}] = 0 \qquad \text{almost surely.} \tag{25}$$

Auto-Calibration

Suppose the conditional distributions $F_{Y|m(\boldsymbol{X})=z}$ are in $\mathcal{F}$ for almost all $z \in \mathbb{R}$. Then m is auto-calibrated for T if and only if

$$\mathbb{E}[V(m(\boldsymbol{X}), Y)|m(\boldsymbol{X})] = 0 \qquad \text{almost surely.} \tag{26}$$

Note

By the tower property of the conditional expectation, conditional calibration implies auto-calibration for identifiable functionals with a sufficiently rich class $\mathcal{F}$.

Unconditional Calibration

Definition

Let V be any strict identification function for T. We say that $m(\boldsymbol{X})$ is **unconditionally calibrated** for T relative to V if $\mathbb{E}[V(m(\boldsymbol{X}), Y)] = 0$.

Unless $m(\boldsymbol{X})$ is constant, and in stark contrast to conditional calibration and auto-calibration, the notion of unconditional calibration depends on the choice of the identification function V used.

Assessing Calibration: Overview

❗

Notion	Definition	Check
conditional calibration	$m(\boldsymbol{X}) = T(Y|\boldsymbol{X})$	$\mathbb{E}[V(m(\boldsymbol{X}), Y)|\boldsymbol{X}] = 0$ *a.s.*
auto-calibration	$m(\boldsymbol{X}) = T(Y|m(\boldsymbol{X}))$	$\mathbb{E}[V(m(\boldsymbol{X}), Y)|m(\boldsymbol{X})] = 0$ *a.s.*
unconditional calibration	$\mathbb{E}[V(m(\boldsymbol{X}), Y)] = 0$	$\mathbb{E}[V(m(\boldsymbol{X}), Y)] = 0$

Table: Types of calibration for an identifiable functional T with strict identification function V.

Assessing Calibration: Test Functions

How to measure *good* calibration on a sample level?

Test functions

$$\mathbb{E}[V(m(\boldsymbol{X}), Y)|\boldsymbol{X}] = 0 \quad a.s.$$

is equivalent to

$$\mathbb{E}[\varphi(\boldsymbol{X})V(m(\boldsymbol{X}), Y)] = 0 \quad \text{for all (measurable) test functions } \varphi\colon \mathcal{X} \to \mathbb{R}. \quad (27)$$

Similarly, auto-calibration is equivalent to

$$\mathbb{E}[\varphi(m(\boldsymbol{X}))V(m(\boldsymbol{X}), Y)] = 0 \quad \text{for all (measurable) test functions } \varphi\colon \mathbb{R} \to \mathbb{R}. \quad (28)$$

Here, φ is a univariate function (only).

Practical Considerations

▶ Check for calibration on the training as well as on the test set.
▶ Quantify calibration by making a choice for the test function φ and report $\overline{V}_\varphi(m; D) = \frac{1}{n} \sum_{i=1}^{n} \varphi(\boldsymbol{x}_i) V(m(\boldsymbol{x}_i), y_i)$. Consider at least
  ▶ $\varphi(\boldsymbol{x}) = 1$
  ▶ all projections to single components of the feature vector $\boldsymbol{x}$
  ▶ $\phi(\boldsymbol{x}) = m(\boldsymbol{x})$ to assesses auto-calibration.

  Continuous features can be binned, e.g., $\varphi(\boldsymbol{x}) = \mathbb{1}\{\text{lower} \leq x^1 < \text{upper}\}$.
▶ Assess calibration visually:
  ▶ Plot the generalised residuals $V(m(\boldsymbol{x}_i), y_i)$ versus $\varphi(\boldsymbol{x}_i)$ for the above choices of test functions φ. Average of generalised residuals should be around 0 for all values of the test function.
  ▶ Another possibility is to plot the values of $\overline{V}_\varphi(m; D)$ for different φ, for instance projections to single feature columns.
  ▶ A reliability diagram assesses auto-calibration: a graph of the mapping $m(\boldsymbol{x}) \to T(Y|m(\boldsymbol{x}))$, see [5]. $T(Y|m(\boldsymbol{x}))$ can be estimated by isotonic regression of y_i against $m(x_i)$. For an auto-calibrated model, the graph is the diagonal line.

Unconditional Calibration in Numbers

Would we have made profit or loss (on test set) on the portfolio?
Note: Ideally neither loss nor profit, i.e. *balanced*.
$n_{\text{test}} = 20504$

	$\frac{1}{n} \sum_i m(\mathbf{x}_i) - y_i$	p-value of t-test
Trivial	**−24**	9.5×10^{-1}
GLM Gamma	−1207	8.8×10^{-4}
GLM Poisson	125	7.3×10^{-1}
XGBoost	−2044	1.4×10^{-8}

$\Rightarrow$ **unconditional calibration:** $\mathbb{E}[m(\mathbf{X}) - Y] \approx 0$

Unconditional Calibration in Numbers

Would we have made profit or loss (on test set) on the portfolio?
Note: Ideally neither loss nor profit, i.e. *balanced*.
$n_{\text{test}} = 20504$

	$\frac{1}{n}\sum_i m(\mathbf{x}_i) - y_i$	p-value of t-test
Trivial	-24	9.5×10^{-1}
GLM Gamma	-1207	8.8×10^{-4}
GLM Poisson	125	7.3×10^{-1}
XGBoost	-2044	1.4×10^{-8}
XGBoost corr	96	7.9×10^{-1}

Recalibrate XGBoost by a multiplicative constant (on training set).

$\Rightarrow$ **unconditional calibration:** $\mathbb{E}[m(\mathbf{X}) - Y] \approx 0$

Calibration Conditional on Gender

Is there a gender bias in the models?

Mean bias by Gender (test set)
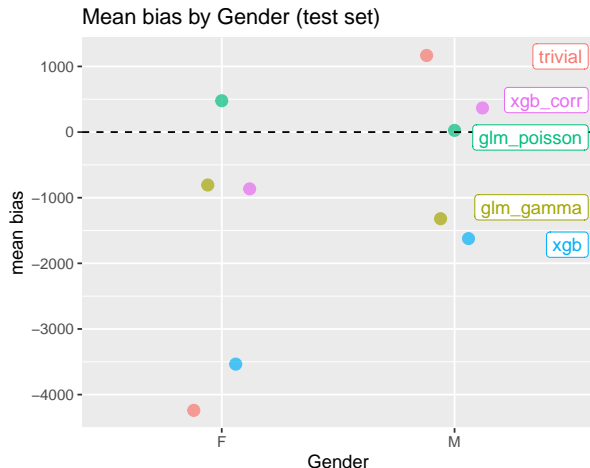
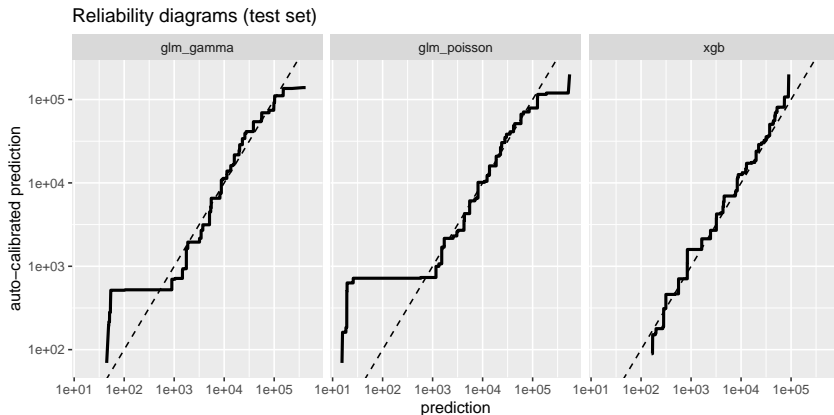| model	$\frac{1}{n} \sum_{i \in subset} m(\boldsymbol{x})_i - y_i$ bias F	bias M
Trivial	-4240	1167
GLM Gamma	-807	-1320
GLM Poisson	**477**	**26**
XGBoost	-3536	-1623
XGBoost corr	-865	367

$\Rightarrow$ **conditional calibration:**
$\mathbb{E}[m(\boldsymbol{X}) - Y | \boldsymbol{X}] \approx 0$

Auto-Calibration

Are policies with same (actuarial) price self-financing?

Reliability diagram: Estimate $\mathbb{E}[Y|m(\boldsymbol{X})]$ via isotonic regression (PAV) and plot vs $m(\boldsymbol{X})$.

Reliability diagrams (test set)

$\Rightarrow$ **auto-calibration:** $\mathbb{E}[m(\boldsymbol{X}) - Y|m(\boldsymbol{X})] \approx 0$

Exercises

Exercise 15

Find a strict identification function $V(z_1, z_2, y) : \mathbb{R}^3 \to \mathbb{R}^2$ for the pair (mean, variance).

Exercise 16

Plot a reliability diagram for your model(s). Hint: The R package monotone might help.

Exercise 17

What are the differences between assessing calibration and the analysis of residuals with ordinary least squares.

Table of Contents

Probabilistic Binary Classification

Probabilistic binary classification

▶ Dichotomous outcome space $O = \{0, 1\}$, $Y \in O$

▶ Probability equals the expectation: $p = \mathbb{P}(Y = 1|\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$

Consequences

▶ Point prediction of the expectation is a fully probabilistic prediction.

▶ Scoring rules and scoring functions coincide.

▶ Proper scoring rules for binary classification are consistent scoring functions for the expectation and are given by the Bregman functions in Eq. 13.
Log loss and Brier score (squared error) are the most used once.

▶ Identification function is $V(z, y) = z - y$.

▶ Prefer probabilistic classifiers (predict p) over deterministic ones (predict 0 or 1).
$\Rightarrow$ More informative predictions and deliberate choice of threshold(s) c:
$m(X) \approx \mathbb{P}(Y = 1|\boldsymbol{X}) \geq c \Rightarrow$ decide for best action assuming $Y = 1$.

Reliability Diagram and Score Decomposition

Reliability diagrams (test set)
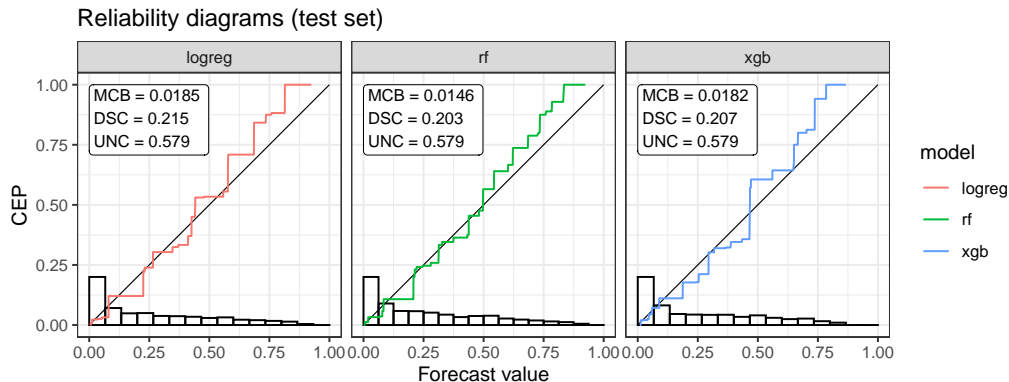
Figure: telco customer churn data set (openml id 42178) with log loss decomposition.

Binary Decision-Making

- Observation domain $O = \{0, 1\}$
- Action domain A comprises potential actions of a decision maker.
- Assume 2 possible actions $A = \{a_0, a_1\}$.

Cost matrix for a decision maker ❗

Cost of action a given outcome Y with reasonableness assumption $c_{ij} > c_{jj}$.

		observed	
		$Y = 0$	$Y = 1$
action	a_0	c_{00}	c_{01}
	a_1	c_{10}	c_{11}

Same decision when adding constants column-wise or multiplying all entries by a constant.

		observed	
		$Y = 0$	$Y = 1$
action	a_0	0	c_1
	a_1	c_0	0

Decision rule

Given information $\boldsymbol{X}$, a decision maker takes action $m(\boldsymbol{X}) = a \in A = \{0, 1\}$ ($a_0 = 0, a_1 = 1$). Minimising the expected cost she decides for a_1 when

$$p c_{01} + (1 - p)c_{00} > p c_{11} + (1 - p)c_{10} \Rightarrow p > \frac{c_{10} - c_{00}}{c_{10} - c_{00} + c_{01} - c_{11}} = \frac{c_0}{c_0 - c_1}.$$

Deterministic Classification: Decision Rule

Deterministic classifier
- $m(\boldsymbol{X}) \in \{0, 1\}$
- Target functional $T \in \{0, 1\}$

Cost-weighted misclassification error

Consider a decision maker:
- Cost c_0 for false positives (predict $m(\boldsymbol{X}) = 1$ while $Y = 0$ materialises)
- Cost c_1 for false negative ($m(\boldsymbol{X}) = 0$ while $Y = 1$)

With cost ration $c = \frac{c_0}{c_0 + c_1}$, $z, y \in \{0, 1\}$, this gives:

$$S_c(z, y) = (1 - c) \underbrace{y(1 - z)}_{\text{false neg}} + c \underbrace{(1 - y)z}_{\text{false pos}} = (\mathbb{1}\{z \geq y\} - (1 - c))(z - y) \qquad (29)$$

This turns out to be the pinball loss for the α-quantile with $\alpha = 1 - c$.

Bayes Classifier

Optimal predictions, aka Bayes classifier, are α-quantiles of $\mathbb{P}(Y|\boldsymbol{X})$ with $\alpha = 1 - c$:

$$m^\star(x) = \begin{cases} 0, & \mathbb{P}(Y = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}) < c \\ 1, & \mathbb{P}(Y = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}) > c \\ 0 \text{ or } 1, & \mathbb{P}(Y = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}) = c \end{cases} \tag{30}$$

In terms of probabilistic predictions $z \in [0, 1]$ with explicit thresholding, S_c reads

$$S_c(z, y) = y(1 - c) \cdot \mathbb{1}\{z \leq c\} + (1 - y)c \cdot \mathbb{1}\{z > c\}, \quad y \in \{0, 1\}, \ z \in [0, 1].$$

Special case $c = \frac{1}{2}$

For equal cost for false positives and false negatives, the ideal model is the **median** of $\mathbb{P}(Y|\boldsymbol{X})$. For a dichotomous response variable Y, the **median** equals the **mode**. The only strictly consistent scoring function for the mode—up to equivalence—is given by

$$S(z, y) = 2S_{\frac{1}{2}}(z, y) = \mathbb{1}\{z \neq y\} = |z - y|, \quad z, y \in \{0, 1\},$$

which is the zero-one loss of (16), also known as $1 -$ accuracy.

Confusion Matrix

Confusion matrix / Binary contingency table

		observed	
		$Y = 0$	$Y = 1$
predicted	$m(\boldsymbol{X}) = 0$	true negative (TN)	false negative (FN)
	$m(\boldsymbol{X}) = 1$	false positive (FP)	true positive (TP)

Lots of derived scores

- accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- hit rate HR $= \frac{TP}{TP+FN}$ (sensitivity, recall, true positive rate)
  population level $\mathbb{P}(m(\boldsymbol{X}) = 1 | Y = 1)$
- false alarm rate FAR $= \frac{FP}{TN+FP}$ (fall-out, false positive rate)
  population level $\mathbb{P}(m(\boldsymbol{X}) = 1 | Y = 0)$

- true negative rate
- F_1-score
  $\frac{2TP}{2TP+FP+FN}$
- . . .

Miscellaneous

Further scores:

- ▶ (ROC) Area Under the Curve (AUC) and Gini index G with $G = 2AUC - 1$
  AUC is no proper scoring rule, only if one restricts to the class of auto-calibrated models[9].

- ▶ Hinge loss: $S(z, y) = \max(0, 1 - \tilde{y}z)$
  with $\tilde{y} = 2y - 1 \in \{-1, +1\}$ for $z \in [-1, 1]$ or $z \in \mathbb{R}$

Graphical tools:

- ▶ Reliability diagram

- ▶ Receiver operator characteristic (ROC)
  Parametric curve $(FAR(c), HR(c))$ with threshold $c \in [0, 1]$, e.g.,
  $HR(c) = \mathbb{P}(m(\boldsymbol{X}) > c \mid Y = 1)$.

- ▶ Cumulative Accuracy Profile (CAP)[9]

---

[9]   C. Lorentzen, M. Mayer and M.V. Wüthrich. "Gini Index and Friends" (October 14, 2022). doi:10.2139/ssrn.4248143
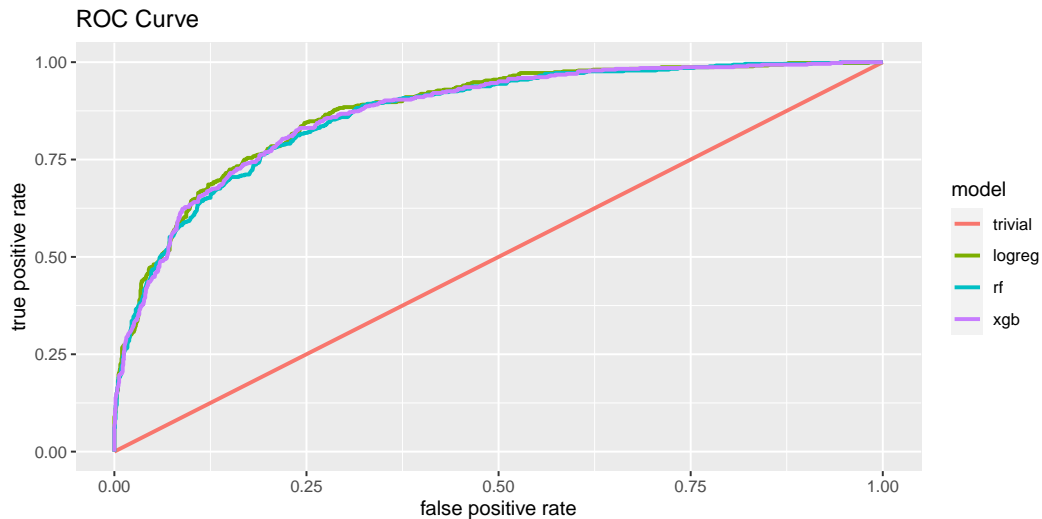
Receiver Operator Characteristic Curve

Figure: telco customer churn data set (openml id 42178).

Imbalanced Classes

Imbalanced Classes

One class significantly outnumbers the other one. Typically, main interest is in the rare class $Y = 1$, called minority class, with $P(Y = 1) \ll P(Y = 0)$, e.g., with a ratio of 1:10, 1:1000 or even more unbalanced.

Note

▶ Foremost, the problem with imbalanced classes is the little information content. Example: How much information is added by one more coin toss with a probability p for $Y = 1 = $ "head"? Every toss adds Shannon entropy $H = -p \log_2 p - (1 - p) \log_2(1 - p)$. H is maximised for $p = 1/2$ giving $H = 1$ bit of information and minimised by $p \in \{0, 1\}$ giving $H = 0$ bits. **Data points for imbalanced classes therefore have little information content.**

▶ Despite the ubiquitous met over- and undersampling techniques: There is no "imbalanced class problem" for well calibrated probabilistic classifiers like Logistic regression—other than the little available information content.

▶ **Only remedy: More data**

Exercises

Exercise 18

Derive the Bayes classifier of Eq. (30) from the cost-weighted misclassification error.

Exercise 19

Derive the Bayes classifier for the Hinge loss.

Exercise 20

Plot a ROC curve for your model(s).

Table of Contents

Bibliographie I

Books on responsible ML or AI

▶ Alyssa Simpson Rochwerger and Wilson Pang. **Real World AI: A Practical Guide for Responsible Machine Learning**. Lioncrest Publishing, 2021

▶ Patrick Hall, James Curtis, and Parul Pandey. **Machine Learning for High-Risk Applications**. O'Reilly Media, Inc., 2022

Model evaluation & scoring functions

▶ Tobias Fissler, Christian Lorentzen, and Michael Mayer. "Model Comparison and Calibration Assessment: User Guide for Consistent Scoring Functions in Machine Learning and Actuarial Practice". In: (2022). DOI: 10.48550/ARXIV.2202.12780

▶ T. Gneiting. "Making and Evaluating Point Forecasts". In: **Journal of the American Statistical Association** 106.494 (2011), pp. 746–762. DOI: 10.1198/jasa.2011.r10138. arXiv: 0912.0902 [math]

Scoring rules

Bibliographie II

▶ T. Gneiting and A. E. Raftery. "Strictly Proper Scoring Rules, Prediction, and Estimation". In: **Journal of the American Statistical Association** 102 (2007), pp. 359–378. DOI: 10.1198/016214506000001437. URL: http://www.stat.washington.edu/people/raftery/Research/PDF/Gneiting2007jasa.pdf

▶ A. Buja, W. Stuetzle, and Y. Shen. **Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications**. Tech. rep. University of Pennsylvania, 2005. URL: http://www-stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf

Explainability

▶ C. Lorentzen and M. Mayer. "Peeking into the Black Box: An Actuarial Case Study for Interpretable Machine Learning". In: **SSRN Manuscript ID 3595944** (2020). DOI: 10.2139/ssrn.3595944.

▶ Christoph Molnar. **Interpretable Machine Learning**. 1st ed. Raleigh, North Carolina: Lulu.com, 2019. ISBN: 978-0-244-76852-2. URL: https://christophm.github.io/interpretable-ml-book