Binäres / Dichotomes Merkmal y, ein oder mehrere (mindestens) nominalskalierte Merkmale X
(nur zwei Kategorien)

Setting: $y \in \{0, 1\}$ binary target variable; feature vector x
often $y=1$ as "positive case", $y=0$ as "negative case"

Ziel: Vorhersage/Diagnose $\hat{y}^{(i)}$ based on score/probability $f(x^{(i)})$ with some threshold c (scalar or vector) e.g. using Logistic regression
Prognosegüte [Performance Measure] gesucht für diese binary classification task

Confusion Matrix

		True Class: y pos. (1)	True Class: y neg. (0)	
Predicted Class	$\hat{y}$ pos.	True Positive (TP)	False Positive (FP)	# pos. pred.
	$\hat{y}$ neg.	False Negative (FN)	True Negative (TN)	# neg. pred.
		# positive cases (n_+)	# negative cases (n_-)	n

$\mathbb{P}(y=1)$: Prevalence $= \frac{\text{\# positive cases}}{\text{Total population } (n)}$

ROC-metrics

Sensitivity / Recall
True positive Rate (TPR): $\rho_{TPR} = \frac{TP}{TP+FN}$
→ Proportion of TP among pos. cases $\mathbb{P}(\hat{y}=1 \mid y=1)$

Specifity
True Negative Rate (TNR): $\rho_{TNR} = \frac{TN}{TN+FP}$
→ Proportion of TN among neg. cases $\mathbb{P}(\hat{y}=0 \mid y=0)$

Precision $\mathbb{P}(y=1 \mid \hat{y}=1)$
Positive Predictive Value: $PPV = \frac{TP}{TP+FP}$ If we predict $\hat{y}=1$, how likely is it a true 1? → Proportion of "real" (correctly class.) pos. among all pos. pred.
(jeweils 1 − ... für False-Anteil)

Negative Predictive Value: $NPV = \frac{TN}{TN+FN}$ If we predict $\hat{y}=0$, how likely is it true 0? → Proportion of correctly classified neg. among all neg. pred.

$ACC = \mathbb{P}(\hat{y}=y)$ $ACC = \mathbb{P}(\hat{y}=1, y=1) + \mathbb{P}(\hat{y}=0, y=0)$

Accuracy (ACC): $\rho_{ACC} = \frac{TP+TN}{n}$ Proportion of correct predictions $ACC = \underbrace{\mathbb{P}(\hat{y}=1 \mid y=1)}_{\text{Sensitivity}} \cdot \underbrace{\mathbb{P}(y=1)}_{\text{prevalence}} + \underbrace{\mathbb{P}(\hat{y}=0 \mid y=0)}_{\text{Specifity}} \cdot \underbrace{\mathbb{P}(y=0)}_{\text{1-prevalence}}$ Fixed weight of Sens & Spec.
Bad metric for small prevalence "Accuracy Paradox" → use e.g. F1-score instead for highly class-imbalance

Fall-out
False Positive Rate: $FPR = \frac{FP}{TN+FP}$ } # neg. cases $= 1 - TNR$

Miss Rate
False Negative Rate: $FNR = \frac{FN}{TP+FN}$ } # pos. cases $= 1 - TPR$

Dont confuse with False Discovery Rate: $FDR = FP / (TP+FP) = 1 - PPV$

F1-score: Its very difficult to achieve high **PPV** and high **TPR** simultaneosly
A classifier predicting more positives will be more sensitive (higher TPR), but will also tend to give more false positives (lower TNR, lower PPV)
A classifier predicting more negatives will be more precise (higher PPV), but will also produce more false negatives (lower TPR)

F1 score aims to maximize PPV and TPR
Balances them, but leans towards the smaller value
⇒ max. $\rho_{F1} \in [0, 1]$ best case: $PPV = TPR = 1$ (no FP, FN) ⤳ $F1 = 1$

$\rho_{F1} = \frac{1}{\frac{1}{2}\left(\frac{1}{PPV} + \frac{1}{TPR}\right)}$ harmonic mean of precision & recall $= 2\,\frac{PPV \cdot TPR}{PPV + TPR} = \frac{2TP}{2TP + FN + FP}$ doesnt account for number of True Negatives!

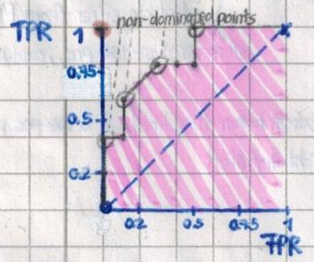
by convention (also for both = 0 (or NA)): PPV or $TPR = 0$: $F1 = 0$ Always predict neg: $F1 = 0$
Always predict pos: $F1 = 2 \cdot PPV / (PPV + 1) = 2 \cdot n_+ / (n_+ + n)$ ($TPR = 1$) big for big n_+, small for small n_+

Indicates Reliability of Predictions $\hat{y}=1$

ROC Curve

(Sensitivity / Recall, 1 − Specifity)
Insensitive to scaling of true class distr. ratio (test! obv. not true for training)
Interesting: FPR, TPR stay the same if we scale ratio n_+/n_- at predict. time

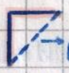
Comparing classifiers by plotting (TPR, FPR) values for all possible thresholds c, want max. TPR & min. FPR
could also use two other trade-off metrics such as (TPR, PPV) where we want to max both



ROC curve invariant to monotone transf. of scores
$h(x) = [\pi(x) \geq c]$ ... $c = 0$, always end in (1,1) & all pos pred.

1.) Rank test observations on decreasing score/prob.
2.) Start with $c = 1$ for probabilistic classif., so we start in (0,0) & no pos. predictions
3.) Iterate through all possible thresholds $c\downarrow$, discrete steps in between one or several obs. with same score
If corresponding true $y = 1$ (pos.): Move TPR up by $1/n_+$ as we have one TP point more
If corresponding true $y = 0$ (neg.): Move FPR right by $1/n_-$ as we have one FP point more
We get diagonal step if, for certain score, we have more than one obs. and they have different y ⇒ we add TP, FP points at the same time

- The closer the ROC-curve is to the left axis with ideal dream point (TPR = 1, FPR = 0) (top left corner) the better our classifier is



ROC of "perfect" classifier, no FP! (until c below all pos.)

- Diagonal represents worst possible classifier that produces random label predictions e.g. 20% pos, 80% neg
↳ random (non discriminant) prediction means $TP = FP$, $FN = TN$ ⇒ $TPR = FPR$ [AUC = 0.5]

all thresholds $c \in (0,1)$ perfectly separate y-groups, AUC = 1
↳ Scorewise ranks all pos. obs $y^{(i)}$ > neg. obs. $y^{(j)}$

Curve below diagonal is impossible. We can then just flip predicted labels (0 → 1 and 1 → 0) $T(F)PR_{new} = 1 - T(F)PR$ which reflects curve at diagonal. Curve below diag. is such an awful classifier that we would never implement it. Flipping makes it so better

Probabilistic classif.: Average predicted prob. should be ≈ prevalence (well calibrated model (captures preval.)) ⇒ pred. prob. estimate real risk/chance

Finding best threshold for specific classifier? Small thresholds (ROC-points in top right corner) will very liberally predict pos. class ⇒ higher TPR, but also (potentially) higher FPR
Big Thresholds (ROC-points in bottom left corner) will very conservatively predict pos. class ⇒ lower FPR, however lower TPR as well

If class imbalances & FP vs. FN costs are unknown, which is often the case, we need to decide threshold manually by visual cue
⇒ identify non-dominated points, assess their TPR/FPR combo, decide which combo is best for task (domain knowledge!) and choose respective threshold

AUC Area under the ROC-curve
treats Sensitivity & Specifity (1 − ...) equally and ignores PPV, NPV!
$AUC \in [0.5, 1]$ (technically 0 possible) Larger AUC ≙ better classifier

Rank-based interpretation: Fraction of all pairs ($y^{(i)}$ pos., $y^{(j)}$ neg.) where $f(x^{(i)}) > f(x^{(j)})$
concordant pairs $(y^{(i)}\text{pos}, f(x^{(i)})), (y^{(j)}\text{neg}, f(x^{(j)}))$ if ✓
$AUC = \frac{N_c + N_e / 2}{N} = \frac{N_c}{N_{\text{unique } y}}$ (Nc: concordant Paare, Ne: Paare mit Ties in y; pair 0,1)
"P(score(random pos. $y^{(i)}$) > score(random neg. $y^{(j)}$))" "Probability of ranking random obs. correctly"

AUC integrates over all feasible thresholds bzw. corresponding points (TPR, FPR) $AUC = \int_{t=0}^{1} ROC_f(t)\,dt$, empirical $\widehat{AUC} = \frac{N_c}{N_{uniquey}}$

Partial AUC

If certain value for TPR or FPR is no acceptable, then it might be useful to fix TPR or FPR to required value minimum and optimize the other metric based on that constraint. Focus only on relevant parts of ROC-curve for our use-case!
↓ Integrate over region after cut-off

- Example: $FPR > 0.2$ is not acceptable
  ⇒ cut out region $FPR > 0.2$ (vertical $\square$ region remains), focus on $FPR \leq 0.2$

- Or $TPR < 0.8$ is so awful that we would never use a classifier in that region.

Rescale Partial AUC metric so that not only does it normalize to $[0,1]$ but also make random labeling correspond to $AUC_{partial} = 0.5$

$$f_{AUC,corrected} = \frac{1}{2}\left(1 + \frac{AUC - AUC_{min}}{AUC_{max} - AUC_{min}}\right)$$ with $AUC_{min} \hat{=}$ Area under diagonal in constrained region and $AUC_{max} \hat{=}$ Area of constrained region

- We could even put constraints on both TPR, FPR and calculate the remaining small area (2 Way partial AUC)

Multiclass AUC

ROC-curves and therefore AUC are only defined for binary classification tasks

Define conditional $AUC(k \mid l)$ where k is positive and l is negative class
↳ For computation subset rows with $y=k$ or $y=l$ from dataset

Careful: Unlike binary case where $AUC(1 \mid 0) = AUC(0 \mid 1)$ it is $AUC(k \mid l) \neq AUC(l \mid k)$!
E.g. with probability $1 - \pi_1 = \pi_0$ but with three or more classes that is generally not true

⇒ using 1-vs-1 scheme: $AUC_{MC} = \frac{1}{g(g-1)} \sum_{k \neq l} AUC(k \mid l) \in [0,1]$

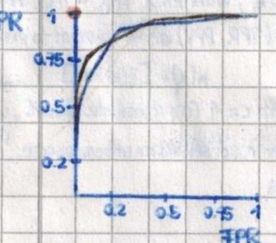
Sidenote] There are other schemes like 1-vs-rest where we only have to average over g AUC values, $AUC(1 \mid Rest)$ ...... $AUC(g \mid Rest)$
This is computationally easier, but create imbalanced classes by design even if original classes were rel. balanced

Precision-Recall-Curves

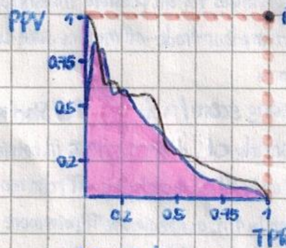
Comparing classifiers by plotting (PPV; TPR) instead for all possible thresholds c, want max. both equally weighted
(Precision, Recall; kind of ignoring TN(R) and FPR)

Might be better than ROC for highly class-imbalanced data $n_- \gg n_+$
often then we are less interested in high TNR ⟺ low FPR, but more in high PPV which reacts more sensitive to abs. ΔFP

For fixed given classifier:
ROC curves for very differently balanced classes could look very similar. PR curves change drastically from balanced to imbalanced classes



ROC-plot



PR-plot

ideal dreampoint (PPV=1, TPR=1) in top right corner — ideal classifier from PR perspective: all pos. pred are true pos., capture all TP

$$PPV = \frac{TP}{TP+FP} \qquad TPR = \frac{TP}{TP+FN}$$

PR-Plot startet links oben bei $(1,0)$ $\hat{=}$ extrem hoher threshold s.d. keine pos. pred. erfolgen
PPV dann per Konvention $=1$, weil $0/0$ nicht definiert

For high TPR: PPV way below 0.5 so even worse than a coinflip guess

ROC plot could make us think that we have two good classifiers here (and similar). PR plot shows them to be quite different and actually pretty bad!

Good: Curve dominates in ROC ⟺ Curve dominates in PR

- In practice, if we only compare a few models on a single task, its not a bad idea to plot both curves and observe
- (Partial) PR AUC can be used for tuning