PREDICTIVE MODEL EVALUATION

(= STATISTICAL LEARNING & DECISION THEORY)

COMPARISON, SELECTION

- INTRODUCTION
- EMPIRICAL RISK MINIMIZATION
- BAYESIAN DECISION THEORY
- CROSS-VALIDATION
- LOSS FUNCTIONS

TODAY (2-3h)

LATER THIS WEEK

PARTIALLY

PREREQ:

- PROB. THEORY & NECESSARY MATH
- PRACTICAL EXP. WITH FITTING & SELECTING MODELS

MODEL EVALUATION IS THE MOST IMPORTANT PART OF MODELING!

1. WE HAVE TO DO IT.
2. ANY MISTAKE WILL BE COSTLY.
3. A GOOD EVAL. WILL PROTECT US.

(THINK OF IT AS A MEASURING TOOL)

MODEL EVAL. IS TIGHTLY BOUND TO LEARNING (FITTING MODELS). FOR PRACTICAL PURPOSES WE OFTEN KEEP THEM SEPARATE.

FOR NOW, WE'LL ALSO KEEP THEM SEPARATE AND FURTHER SUBDIVIDE MODEL EVAL. INTO

A. WHAT TO MEASURE?

MISTAKE: CHOOSE THE WRONG THING TO MEASURE.

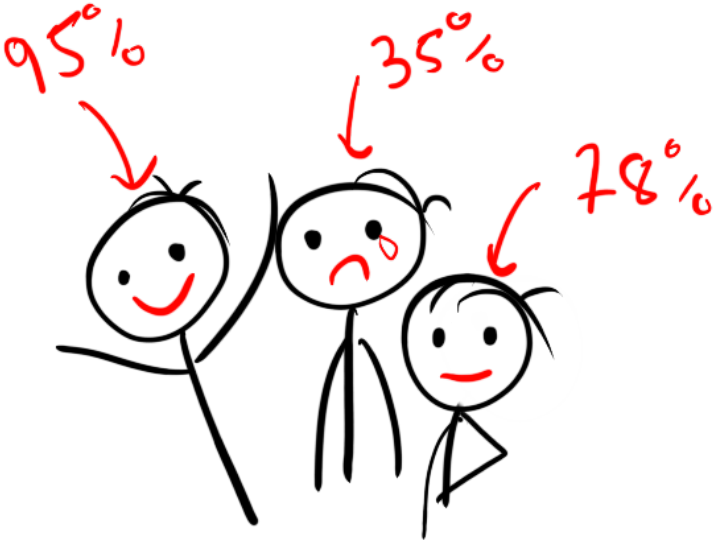
B. HOW TO MEASURE IT?

MISTAKE: MEASURE WITH A LOT OF ERROR. (CHOOSE AN INCORRECT PROCESS)

~~MODEL~~ EVALUATION

STUDENT

- STUDENT = LEARNING ALGO.
- TEACHER = MODEL EVALUATION



(STUDENTS)

A. WHAT TO MEASURE? SOME OPTIONS:

- PERFORMANCE IN ML-RELATED PROBLEMS. SOUNDS REASONABLE
- RELATIVE PERF. IN ML-RELATED PROBLEMS. -11-
- HOW GOOD THE STUDENT'S DANCE MOVES ARE. BAD CHOICE

(B) HOW TO MEASURE IT?

SUPPOSE WE'VE DECIDED ON EVERYTHING, EXCEPT FOR WHO WILL GRADE THE EXAMS:

"GOLD STANDARD"

I. II. III.

MONKEY
- HIGH ERROR
- (HIGH VARIANCE, NO BIAS)

TEACHING ASSISTANT
- PRACTICALLY ACCEPTABLE ERROR
- LOWER VARIANCE
- SOME BIAS (STRICTER, LENIENT)

PROFESSOR

BIAS-VARIANCE DECOMPOSITION (OF AN ESTIMATOR): $E[(\hat{\theta}-\theta)^2] = \underbrace{(E[\hat{\theta}]-\theta)^2}_{\text{BIAS}} + \underbrace{E[(E[\hat{\theta}]-\theta)^2]}_{\text{VARIANCE}}$

OVERFITTING

IN MEASUREMENT IN EDUCATION THERE ARE MANY SOURCES OF BIAS.

THERE IS ONLY 1 THAT WILL BE OF PARTICULAR INTEREST TO US - POSITIVE BIAS DUE TO OVERFITTING.

EXAMPLES:

- STUDENTS LEARN MOSTLY FROM PAST EXAMS
  (IF WE DON'T ACCOUNT FOR THAT, WE'LL OVERESTIMATE)
- EXTREME CASE: EXAM = PRACTICE TEST

WHEN I DONT SHARE EXAMS & KEEP CHANGING PROBLEMS,
I'M NOT TRYING TO BE AN A*****E.
I'M JUST PREVENTING OVERFITTING.

SUMMARY

- DON'T EVALUATE STUDENTS BASED ON THEIR DANCE MOVES, IF YOU WANT THEM TO LEARN MACHINE LEARNING.
- IF YOU WANT THEM TO DANCE, TEACH THEM DANCING.
- DON'T LET A MONKEY GRADE EXAMS.
- DON'T PUT PAST PROBLEMS ON EXAMS.

...AND NOW FOR SOMETHING COMPLETELY DIFFERENT?

MEASURES OF PREDICTIVE PERFORMANCE:

ACCURACY, MEAN SQUARED ERROR (MSE), ROOT MSE, ABSOLUTE ERROR, QUADRATIC SCORE, BRIER SCORE, SPHERICAL SCORE, LOG LOSS, AIC, BIC, WAIC, AUC, SENSITIVITY, SPECIFICITY, F1 ...

LEARNING PARADIGMS:

EMPIRICAL RISK MINIMIZATION, MAXIMUM LIKELIHOOD, BAYESIAN INFERENCE.

VERY OFTEN IN COMBINATION WITH INFORMATION THEORETIC LOSS

1. DO I REALLY NEED ALL THESE MEASURES? AND IF SO HOW DO I CHOOSE?
2. WHY DO PEOPLE OFTEN LEARN MODELS USING ONE, AND THEN EVALUATE THEM USING ANOTHER MEASURE?

LOGISTIC REGRESSION USING MLE

ACCURACY

NOTATION

$\mathcal{X}, \mathcal{Y}$ - INPUT & OUTPUT SPACE

$P_{X,Y}$ - JOINT DISTRIBUTION OF $(X \in \mathcal{X}, Y \in \mathcal{Y})$

$D_n = \{(x_i, y_i)\}_{i=1}^{n}$ - OUR DATASET (ASSUME IID SETTING)

LEARNING ALGORITHM:

LEARNING ALGORITHM A, GIVEN DATASET D_n, PRODUCES MODEL h_n OR $A(D_n) = h_n$

MODEL: $h : \mathcal{X} \to \mathcal{Y}$

HYPOTHESIS SPACE H - SET OF ALL MODELS THAT A CAN PRODUCE.

LOSS FUNCTIONS (WHAT TO MEASURE)

A LOSS FUNCTION ℓ IS A MAP $\ell : Y \times Y \to \mathbb{R}^+$.

EXAMPLES:

- $\ell(y, \hat{y}) = I_{y \neq \hat{y}} = \begin{cases} 0, & \text{if } y = \hat{y} \\ 1, & \text{otherwise} \end{cases}$
- $\ell(y, \hat{y}) = (y - \hat{y})^2$
- $\ell(y, p) = -\log p(y)$

EMPIRICAL RISK MINIMIZATION (ERM)

A FORMAL LEARNING/DECISION-THEORY FRAMEWORK.
(AS ARE MAXIMUM LIKELIHOOD EST. & BAYESIAN INF.)

KEY QUANTITY: MODEL RISK (GENERALIZATION ERROR)

$$R(h) = E_{X,Y}\left[\ell(Y, h(X))\right]$$

MODEL → h
RISK → $R(h)$
DGP → X, Y
TRUE OUTPUT → Y
TRUE INPUT → X
LOSS FUNCTION → ℓ
MODEL PREDICTION → $h(X)$

= EXPECTED LOSS (OVER $P_{X,Y}$). INFORMALLY, WE OFTEN SAY "FOR NEW/UNSEEN DATA".

NOTE: WE CAN RARELY COMPUTE $R(h)$ BECAUSE WE DON'T KNOW $P_{X,Y}$.

EMPIRICAL RISK

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h(x_i))$$

TO DISTINGUISH FROM TRUE RISK

WE KNOW THAT $R_n(h) \xrightarrow{a.s.} R(h)$. WHY?

NOW WE CAN FORMALIZE ERM:

$$h_n = \arg\min_{h \in H} R_n(h)$$

(DON'T WORRY ABOUT OVERFITTING)

AN ERM LEARNING ALG. PRODUCES h_n THAT MINIMIZES EMPIRICAL RISK.

NOTE: h_n IS RANDOM (D_n IS RANDOM) (WE'LL STUDY THE PROCESS, NOT A REALIZATION)

EXAMPLES OF ERM ESTIMATORS:

NOTE: PROOFS IN NOTES

- FOR $Y = \mathbb{R}$ AND QUADRATIC LOSS WE HAVE

$$R_n(\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2$$ (MINIMIZED BY AVERAGE)

CLEAR FROM MLE EST.

- -11- AND ABSOLUTE LOSS WE HAVE

$$R_n(\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}|$$ (MINIMIZED BY MEDIAN)

$\frac{d}{dx}|x| = \text{sign}(x)$

- FOR FINITE Y AND 0-1 LOSS WE HAVE

$$R_n(\hat{y}) = \frac{1}{n}\sum_{i=1}^{n} I_{y_i \neq \hat{y}} = 1 - \frac{1}{n}\sum_{i=1}^{n} I_{y_i = \hat{y}}$$ (MINIMIZED BY MODE)

NEGET MOST POINTS FOR MOST COMMON y_i

- FOR FINITE Y AND LOG LOSS WE HAVE

$$R_n(\hat{p}) = \frac{1}{n}\sum_{i=1}^{n} -\log p(y_i)$$ (MINIMIZED BY REL. FREQ.)

ENCOURAGES US TO MINIMIZE KL DIVERGENCE

ERM WITH LOG LOSS FOR PARAMETRIC MODEL $p(y|\theta)$

$$\theta_{ERM} = \arg\min_{\theta \in \Theta} \left(\frac{1}{n} \sum -\log p(y|\theta) \right)$$

$$= \arg\max_{\theta \in \Theta} \left(\sum \log p(y|\theta) \right)$$

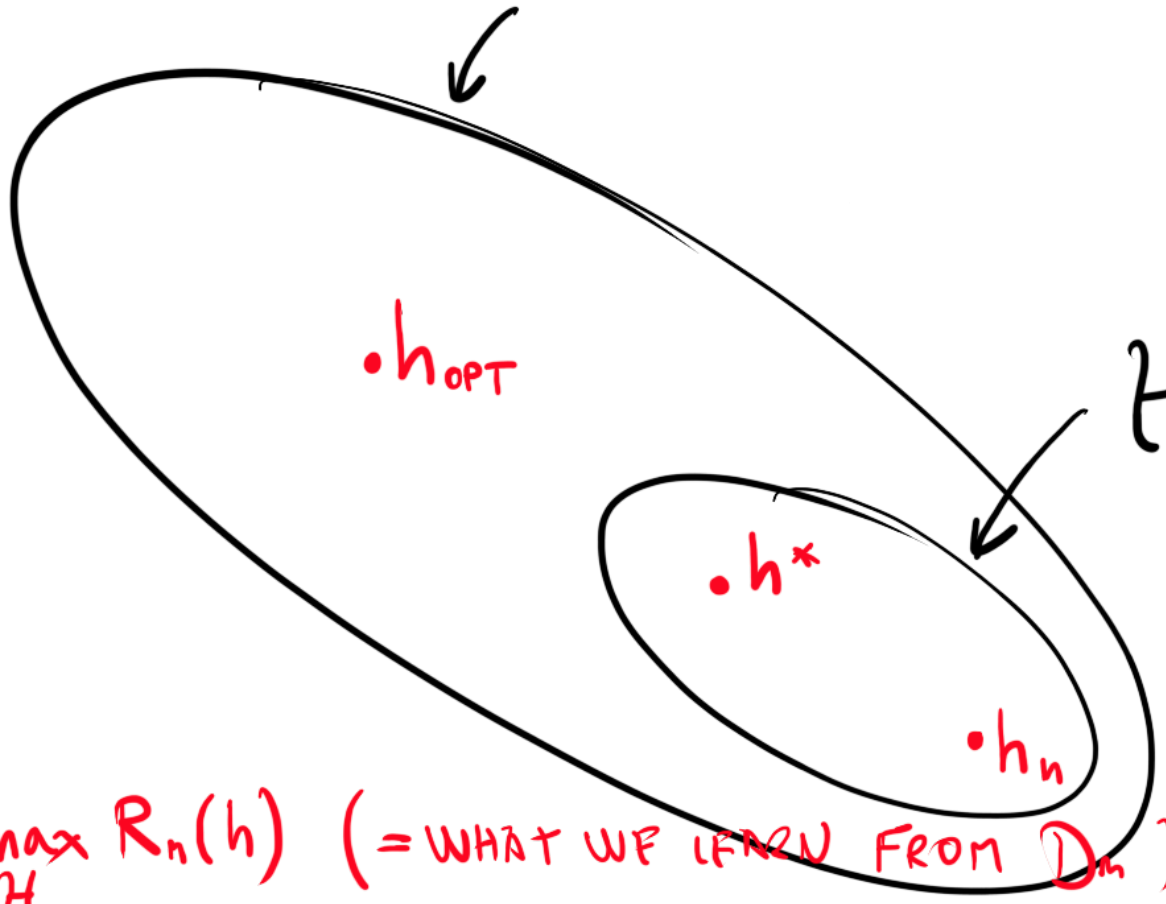
$$= \arg\max_{\theta \in \Theta} \left(\log \prod p(y|\theta) \right)$$

$$= \arg\max_{\theta \in \Theta} \prod p(y|\theta)$$

MAXIMUM LIKELIHOOD ESTIMATION IS A SPECIAL CASE OF ERM!

APPROXIMATION-ESTIMATION DECOMPOSITION

SETUP:



$h_n = \underset{h \in \mathcal{H}}{\arg\max}\, R_n(h)$ (= WHAT WE LEARN FROM D_n)

$h^* = \underset{h \in \mathcal{H}}{\arg\max}\, R(h)$ (= BEST WE CAN HOPE FOR IN $\mathcal{H}$), $R^* = R(h^*)$

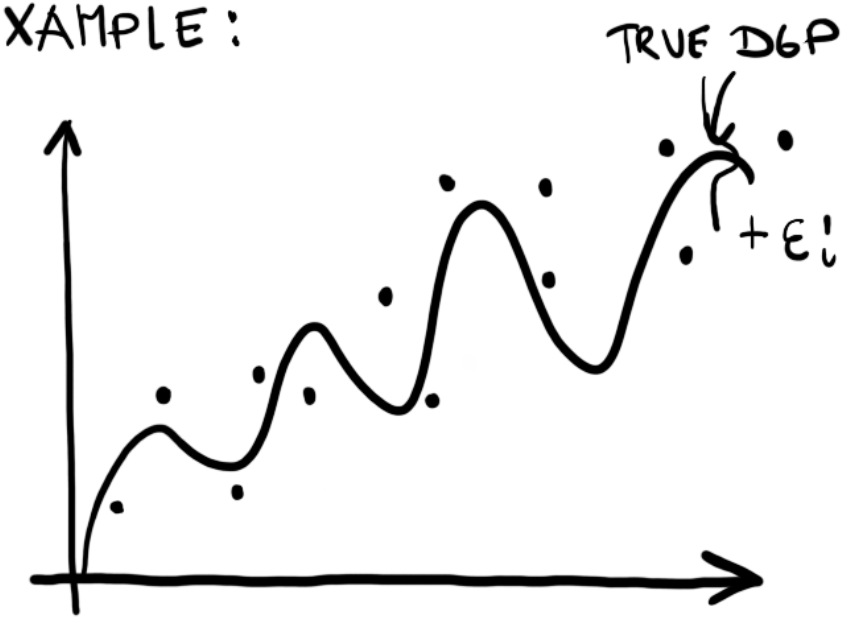
$h_{opt} = \underset{\text{"ALL MODELS"}}{\arg\max}\, R(h)$ (= IDEAL MODEL), $R_{opt} = R(h_{opt})$

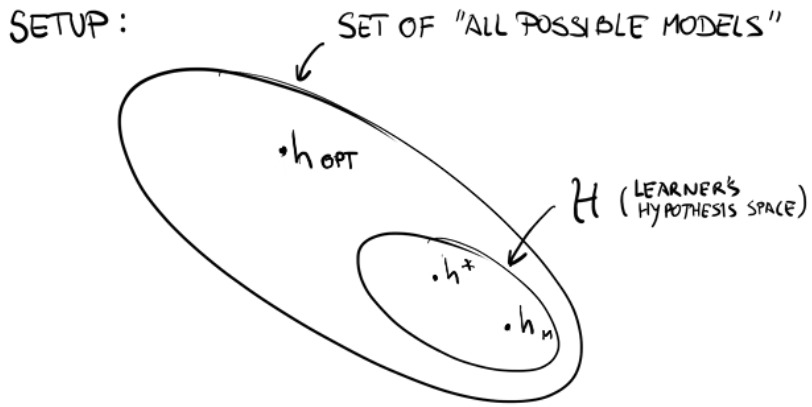
APPROXIMATION-ESTIMATION DECOMPOSITION

$$\underbrace{R(h_n) - R_{OPT}}_{\text{EXCESS RISK}} = \underbrace{(R^* - R_{OPT})}_{\text{APPROXIMATION ERR.}} + \underbrace{(R(h_n) - R^*)}_{\text{ESTIMATION ERR.}}$$

(SORT OF BIAS-VARIANCE DECOMPOSITION)

EXAMPLE:





(POPULAR LEARNING ALG. HAVE GOOD INDUCTIVE BIAS)

CONSISTENCY OF ERM

UNIFORM CONVERGENCE (GLIVENKO-CANTELLI)

IF AND ONLY IF $\sup_{h \in \mathcal{H}} |R_n(h) - R(h)| \xrightarrow{P} 0$,

THEN

$R(h_n) \xrightarrow{P} R(h^*)$ AND $R_n(h_n) \xrightarrow{P} R_n(h^*)$.

THE RISK OF THE MODEL WE LEARN CONVERGES TO RISK OF BEST-CASE MODEL

IN PRACTICE: $\mathcal{H}$ MUSTN'T BE TOO BIG

NOTE: WHY IS $R_n(h) \xrightarrow{a.s.} R(h)$ NOT ENOUGH?

THE PROBLEM IS IF THE CONVERGENCE IS SLOWER FOR SOME h.

UNIFORM CONV. = ALL CONVERGE AT SAME PACE

GENERALIZATION ERROR BOUNDS

TYPICALLY, WE HAVE $R(h_n) > R_n(h_n)$

OR $R(h_n) \leq R_n(h_n) + C$

CAN WE SAY ANYTHING ABOUT C?

IDEA: C DEPENDS ON LEARNER'S CAPACITY TO LEARN!

FORMALIZATION – VC DIMENSION (VAPNIK-CHERVONENKIS)

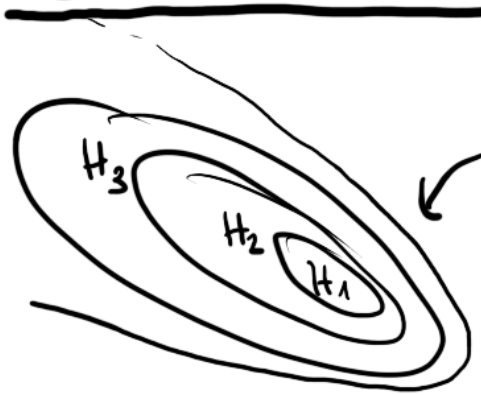
AN EXAMPLE OF A RESULT FROM VC THEORY:

$$R(h_n) \leq R_n(h_n) + O\left(\sqrt{\frac{d_{VC}}{n} \log \frac{n}{d_{VC}} - \frac{1}{n} \log \delta}\right),$$

WITH PROBABILITY $1-\delta$.

(COMPUTING VC-DIMENSION OF LEARNING ALG. IS HARD)

STRUCTURAL RISK MINIMIZATION (SRM)

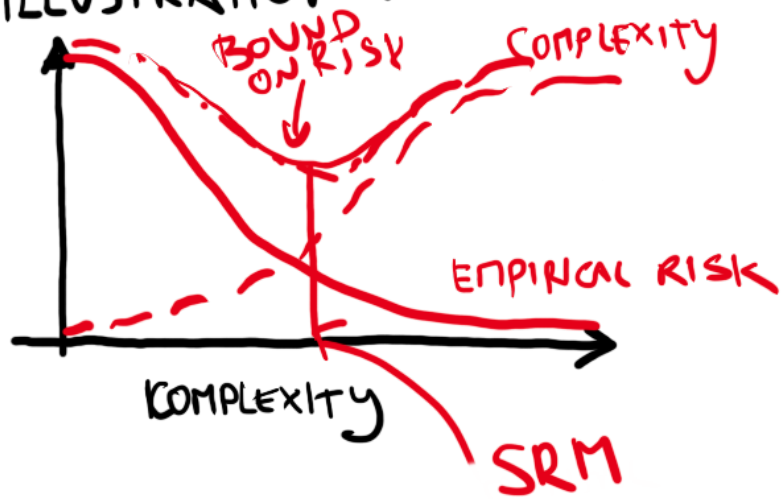


NESTED HYPOTHESIS SPACES $\mathcal{H} \supset \ldots \supset \mathcal{H}_2 \supset \mathcal{H}_1$

SVM: MAXIMIZING MARGIN MINIMIZES BOUND

- FIND BEST $h_n^{(i)}$ FOR EACH $\mathcal{H}_i$
- COMPUTE VC DIMENSION FOR EACH $\mathcal{H}_i$
- PICK h_n THAT MINIMIZES BOUND ON $R(h_n)$

ILLUSTRATION:



$$R(h_n^{(i)}) \leq \underbrace{R_n(h_n^{(i)}) + C(\mathcal{H}_i)}_{\text{MINIMIZE THIS}}$$

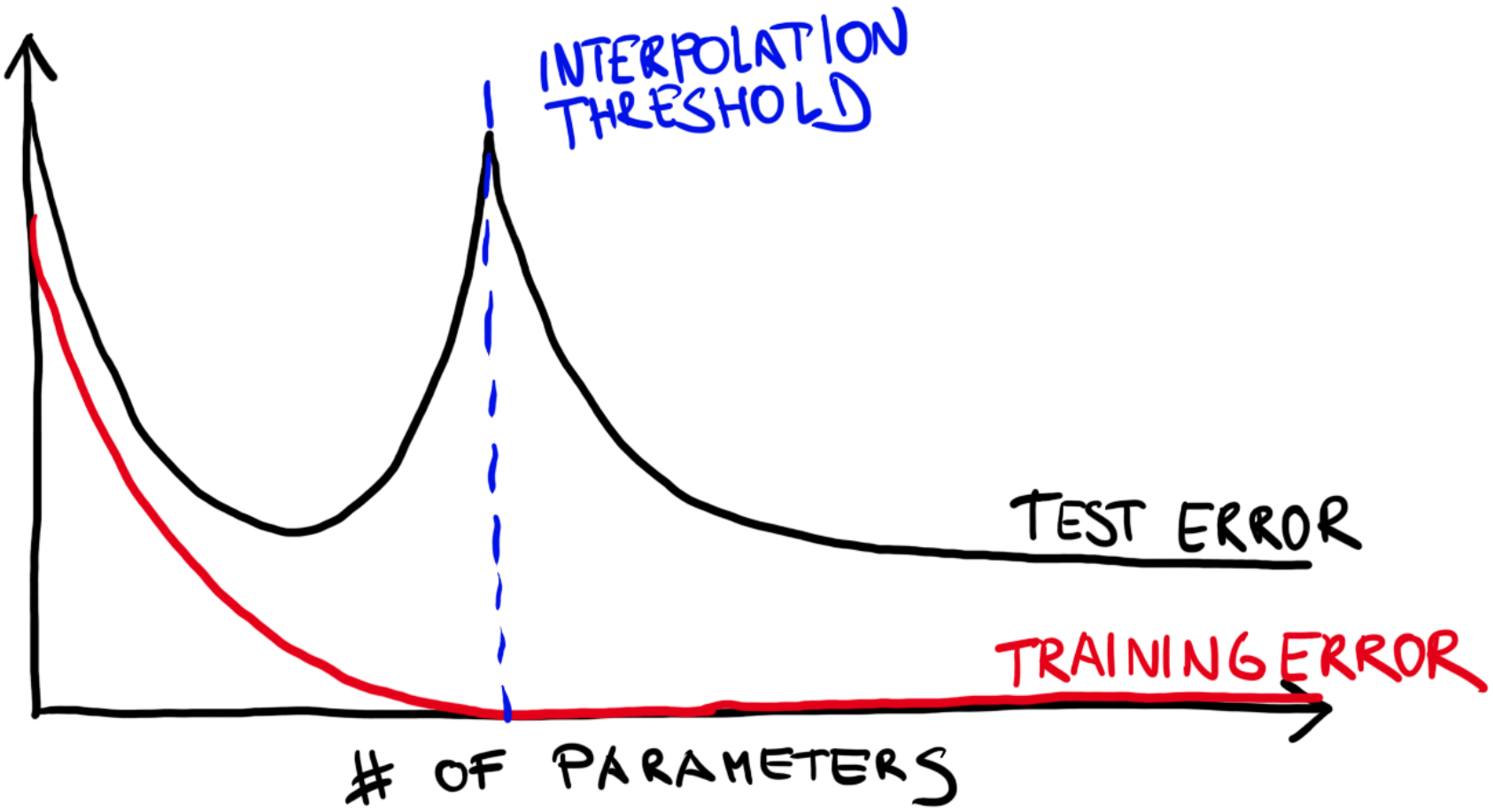
NOTE 1: (REGULARIZATION)

$$h_n = \arg\min_{h \in \mathcal{H}} \left(R_n(h) + \lambda C(h) \right)$$

NOTE 2:

CROSS-VALIDATION INDIRECTLY ESTIMATES MODEL COMPLEXITY (DIRECTLY ESTIMATES TRUE RISK)

DOUBLE-DESCENT



BELKIN ET AL. : "RECONCILING MODERN..."[1]

CHOOSING A LOSS FUNCTION

IN ERM THE DECISION THEORY & LEARNING ARE INSEPARABLE (LOSS FUNCTION IS PART OF LEARNING)

⇒ IT MAKES NO SENSE TO LEARN MODELS WITH ONE LOSS FUNCTION AND THEN COMPARE THEM USING ANOTHER.

(EXCEPTIONS: IF IT'S COMPUTATIONALLY OR OTHERWISE INFEASIBLE, WE CAN USE A SURROGATE LOSS)

ULTIMATELY, IT IS A CASE-BY-CASE CHOICE, BUT WE CAN PROVIDE SOME GUIDELINES:

- DON'T USE A LOSS FUNCTION JUST BECAUSE EVERYONE ELSE IS.
- UNDERSTAND WHAT A LOSS FUNCTION ENCOURAGES.
- IF IN DOUBT, USE MLE/LOG-SCORE (OR, EVEN BETTER, BAYESIAN)

WE'LL DISCUSS THIS NEXT...

SCORING RULES

TYPICALLY SET OF ALL MEASURABLE FUNC. ON $\mathcal{Y}$

A SCORING RULE S IS A MAP $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$.

$= S(p, y)$ TAKES A PROBABILITY/DENSITY AND THE TRUE OUTCOME AND OUTPUTS A SCORE.

(LIKE LOSS FUNCTIONS, BUT FOCUSING ON PMF/PDFs)

EXAMPLES:

$$S(p, y) = \log p(y) \quad \text{(LOG SCORE)}$$

POINT-WISE LOSS FUNCTIONS ARE NOT SCORING RULES, BUT WE CAN MODIFY THEM.

$$S_{MSE}(p, y) = -(E_p[Y] - y)^2 \quad \text{(MSE)}$$

$$S_{0\text{-}1}(p, y) = 1 \text{ (IF } y \text{ IS MODE OF } p\text{), } 0 \text{ OTHERWISE}$$

PROPER SCORING RULES

A SCORING RULE IS **PROPER** IF IT IS MAXIMIZED BY PREDICTING THE TRUE p.

IT IS **STRICTLY PROPER** IF p IS UNIQUE MAX.

SOME STRICTLY PROPER RULES: (FOR DISCRETE ONLY)
LOG SCORE, QUADRATIC (BRIER) SCORE, RANKED PROB. SCORE.

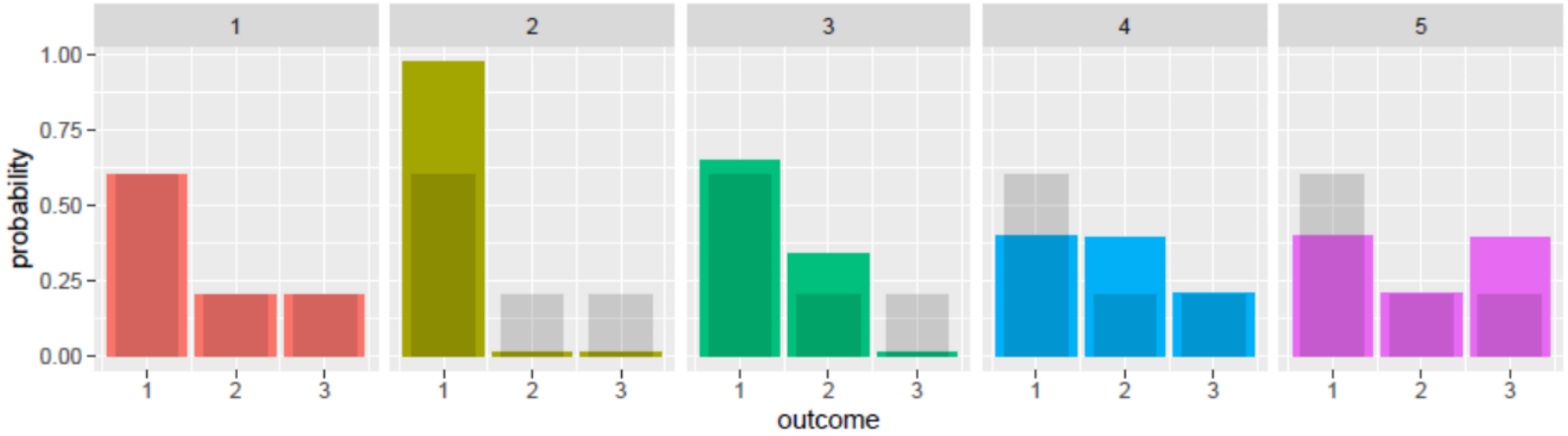
MANY COMMON LOSS FUNCTIONS ARE NOT STRICTLY PROPER!
FOR EXAMPLE: ACCURACY (0-1 SCORE) IS NOT - IT IS MAXIMIZED BY ANY PREDICTION WHOSE MODE IS y!

IN FACT: ANY SCORING RULE THAT INDUCES EQUIVALENCE CLASSES IS NOT STRICTLY PROPER!
(MSE, AUC, ABSOLUTE LOSS, SENS/SPEC., ...)

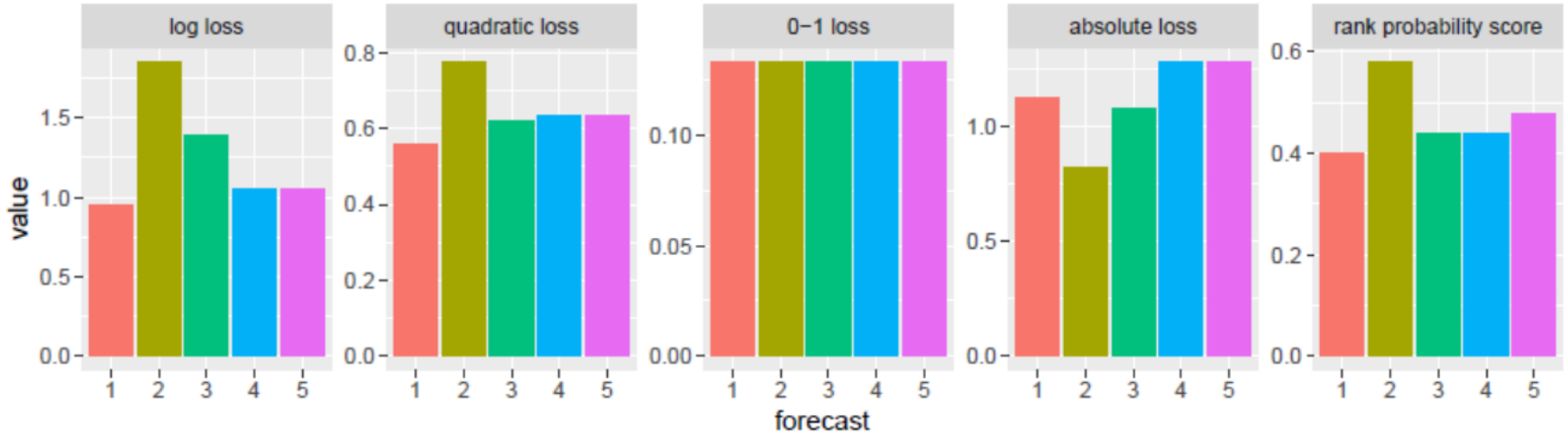
RULES THAT ARE NOT STRICTLY PROPER SHOULD NOT BE USED IF WE WANT TO DO ANY SORT OF INFERENCE!

SOME COMMON LOSS FUNCTIONS ARE NOT EVEN PROPER (ABS. LOSS FOR DISCRETE, AUC).
$\Rightarrow$ THEY ARE NOT MAXIMIZED BY THE TRUE p!

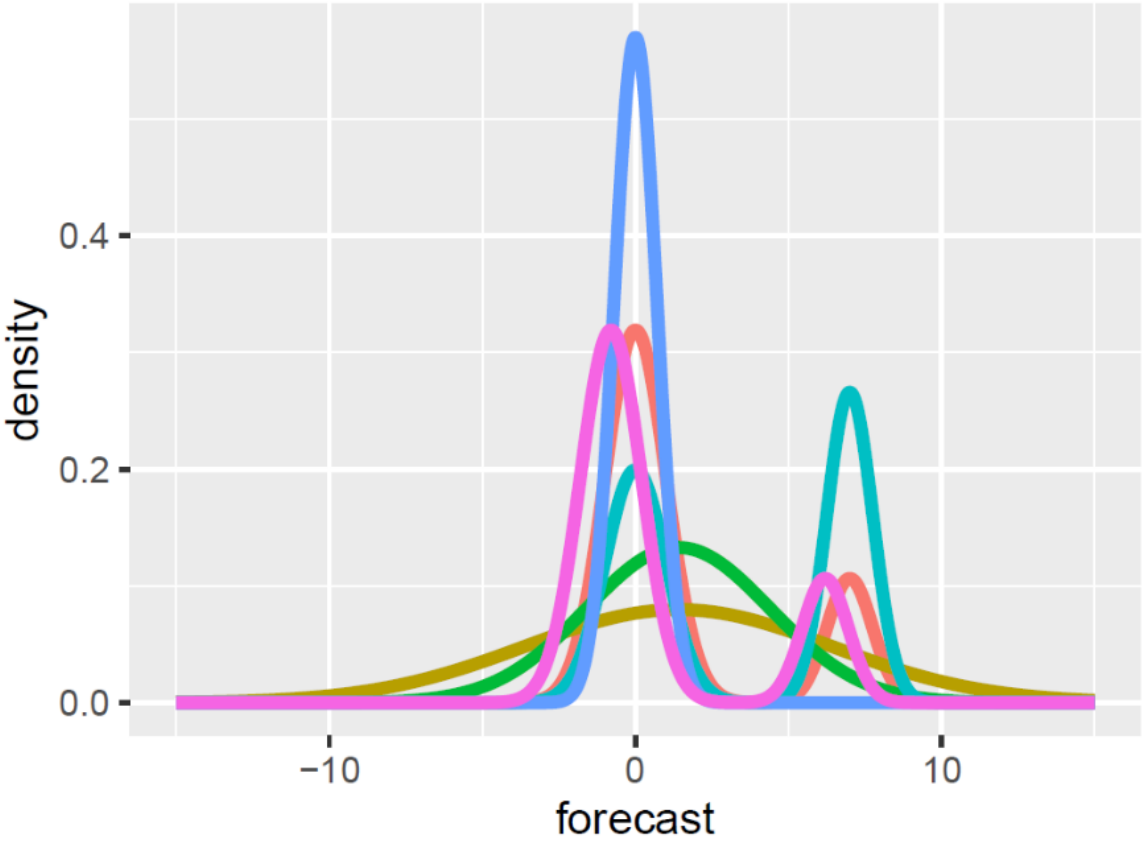
FINALLY, SOME EXAMPLES...



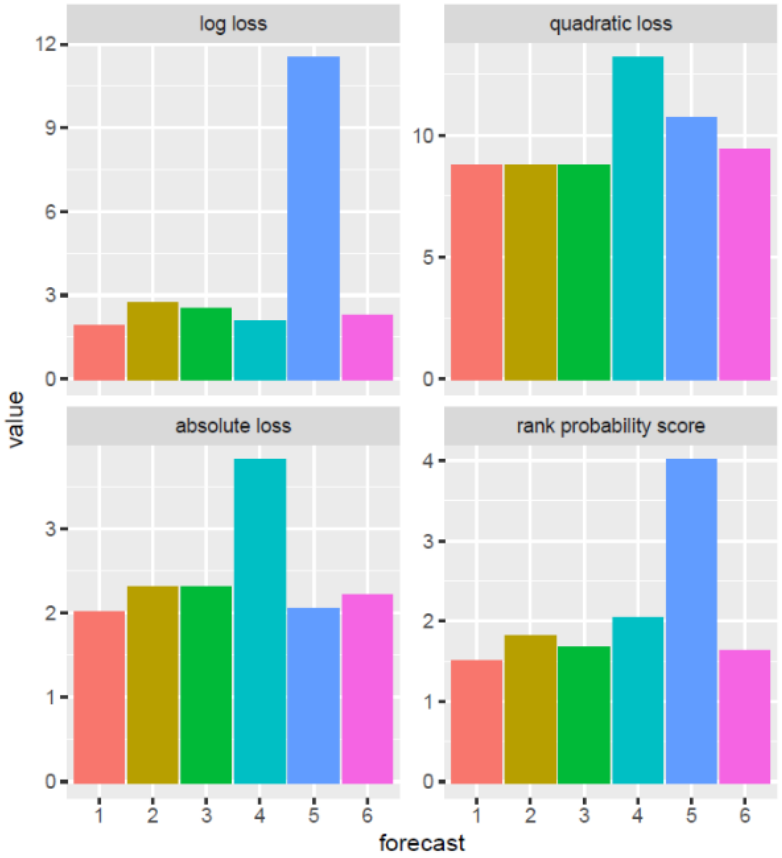
(a) five different probabilistic forecasts (in gray are the true probabilities from 1)



(b) expected loss for five different loss functions



(a) six different density forecasts (true density is in red)



(b) expected loss for four different loss functions