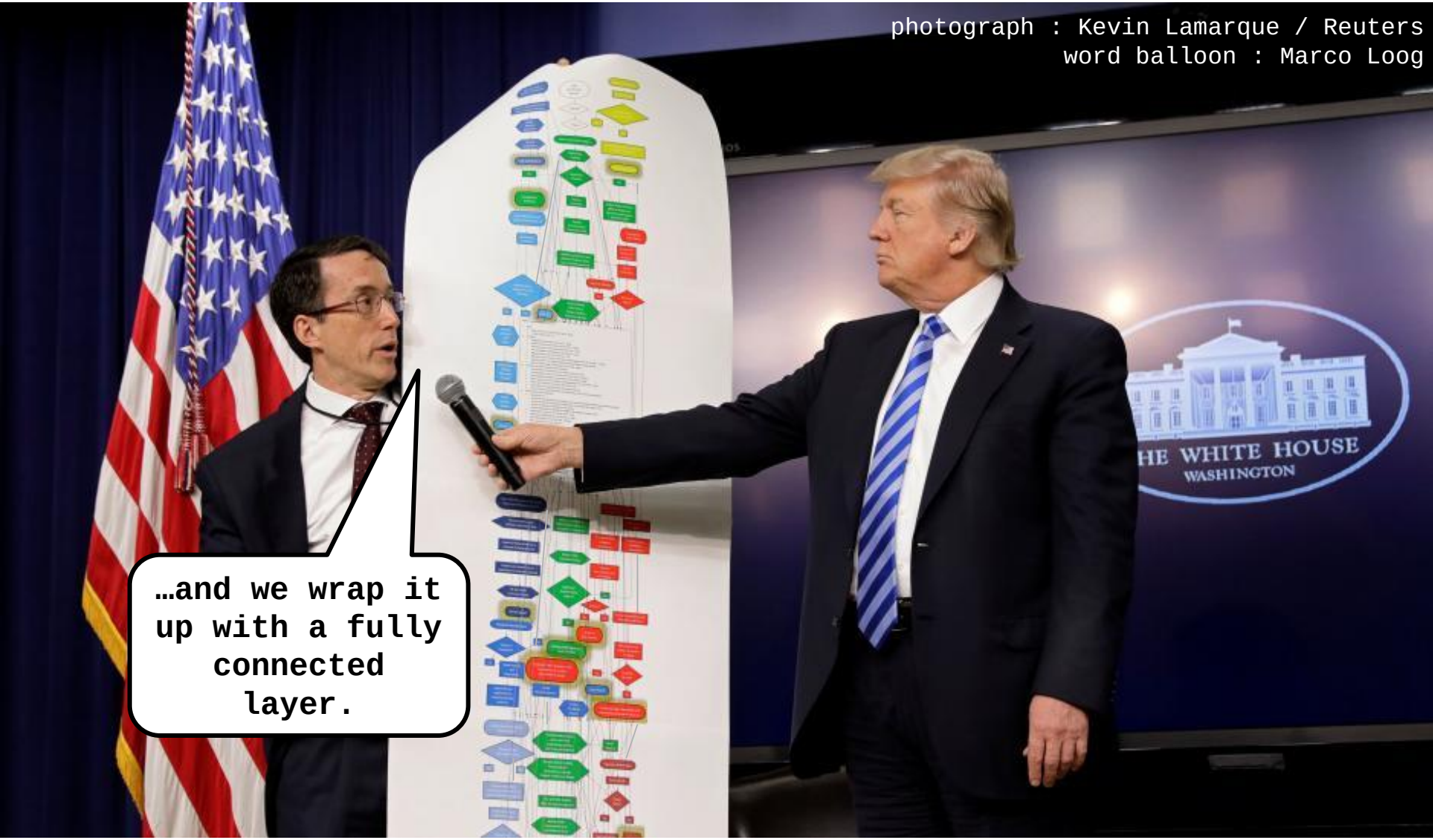

photograph : Kevin Lamarque / Reuters
word balloon : Marco Loog
…and we wrap it up with a fully connected layer.
HE WHITE HOUSE
WASHINGTON

KU / TU Delft

Surrogate Losses
in Classical Machine Learning

Marco Loog

KU / TU Delft

Outline

- Preliminaries and some background
- A basic problem...
- Why objective functions look the way they look
- Semi-supervised learning
- Other learning settings with issues
- Speculate [a bit] on influence of quantum computing

Outline

- Preliminaries and some background
- A basic problem...
- Why objective functions look the way they look
- Semi-supervised learning
- Other learning settings with issues
- Speculate [a bit] on influence of quantum computing
  - Really only a tiny, tiny bit...

Preliminaries

...and Some Background

We Consider

Classical Supervised Learning

- That is, we aim to learn an input-output relation
  - Through generalization from i.i.d. examples from p_{XY}
- Inputs are taken to be d-dimensional vectors
  - E.g. $x \in \mathbb{R}^d$

We Consider

Classification in Particular

- Output is discrete [a class, a category, etc.]
- Basically restrict ourselves to two-class outputs
  - In particular : $y \in \{-1, +1\}$
  - Classifier then is a function : $\mathbb{R}^d \rightarrow \{-1, +1\}$
- Good classifier = classifier that generalizes well
  - Typically measured by [expected] error rate
- Training classifier = estimation of its free parameters
  - Based on some objective and finite number of examples

Generalization

Not Computation

- Focus is largely on non-computational / generalization side of learning
  - It is more about how to formulate learning problem, e.g. what objective function to consider in the first place
  - Not so much about how to perform actual computation, e.g. how to optimize that objective formulated
- But end with speculation on how quantum computing can affect generalization-side of machine learning...

A Bit About Me

[Shameless, I Know...]

- Warning : not hindered by any deep knowledge of QC

- M.Sc. in mathematics, Ph.D. in medical image analysis
- Current focus on concepts and methodologies of machine learning and pattern recognition
  - Slight variations on supervised classification
  - Aim for insight and understanding
  - Research is a combination of theory and empiricism

- I like rather basic problems...

One of Those Basic Problems?

Through a Question to the Audience

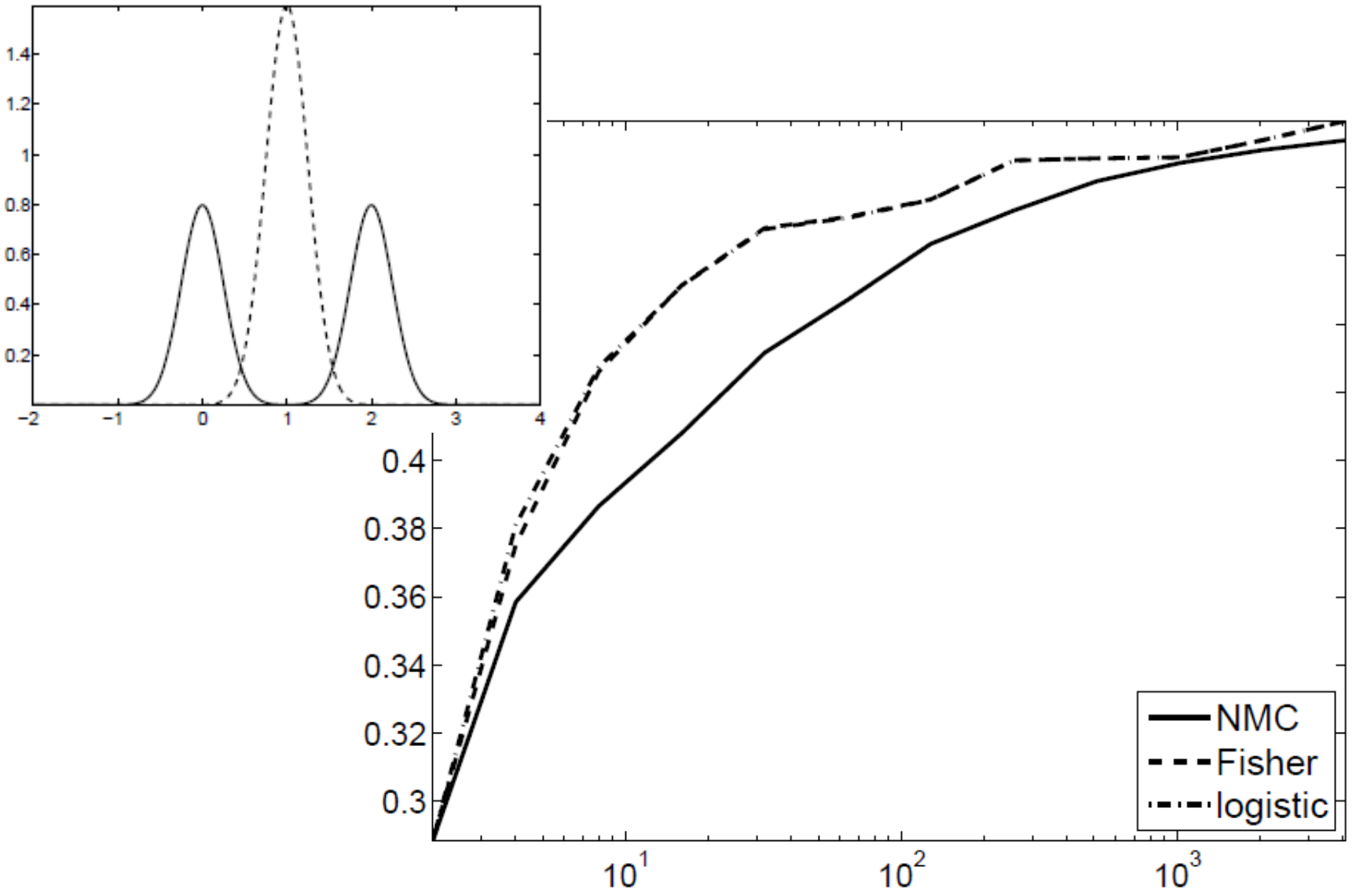
[1,2]

Can You...

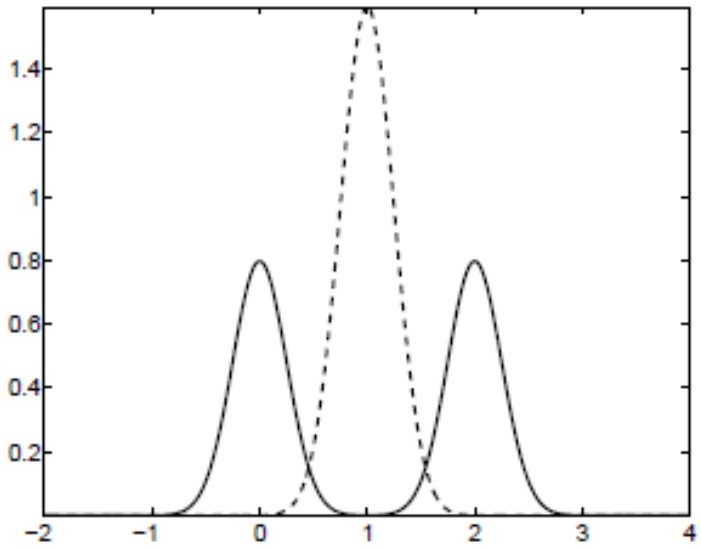
- ...give me a classification problem p_{XY} and a [standard or widely used*] classifier such that expected error rate grows with increasing number of [i.i.d.] training samples?

*Specifications added with thanks to Tamás Kriváchy

E.g. Dipping for Linear Classifiers



More Data ⇏ Better Classifier



- Not even in expectation, i.e., it is not the effect of an unlucky draw
- Part of the reason is the misspecified model
- To fully understand behavior, look at objective functions
- Will also play role in remainder

KU / TU Delft

Objective Functions

Why They Look the Way They Look...

KU / TU Delft

A Technicality

- Rather than considering classifier $f(\,.\,|\theta) : \mathbb{R}^d \to \{-1, +1\}$ directly, one typically considers functions $f(\,.\,|\theta) : \mathbb{R}^d \to \mathbb{R}$
  - From some hypothesis class parametrized by Θ

- Signing then leads to actual classifier : $\operatorname{sign} f(\,.\,|\theta)$
  - E.g. $\theta \in \mathbb{R}^d$ gives a linear classifier with $f(x|\theta) \coloneqq x^T\theta$

What We Actually Want
in Supervised Classification

- Find a [proto]classifier $f(\,.\,|\theta) : \mathbb{R}^d \rightarrow \{-1, +1\}$ such that the expected error is minimized

$$\mathrm{E}([f(x|\theta) \neq y]) \coloneqq$$

$$\int_{\mathbb{R}^d} \sum_{y \in \{-1,+1\}} [\mathrm{sign}\, f(x|\theta) \neq y]\, p_{XY}(x, y)\, dx$$

The First Problem

- True distribution p_{XY} is unknown
  - Merely have a finite [i.i.d.] sample of size N

- Settle for empirical risk

$$\mathrm{E}([f(x|\theta) \neq y]) \approx \frac{1}{N}\sum_{i=1}^{N}[\operatorname{sign} f(x_i|\theta) \neq y_i]$$

The Second Problem

- Still, actual training is typically NP-hard
  - Objective is highly nonconvex in terms of parameters θ

- Settle for easier-to-optimize surrogate loss

$$\mathrm{E}([f(x|\theta) \neq y]) \lessapprox \frac{1}{N}\sum_{i=1}^{N} \ell(f(x_i|\theta), y_i)$$

Surrogate Losses

- Error rate employs 0-1 loss : $[\operatorname{sign} f \neq y]$
- Surrogate gives convex upper bound : $[\operatorname{sign} f \neq y] \leq \ell(f, y)$
- Often consider margin-based losses : $\ell(yf)$

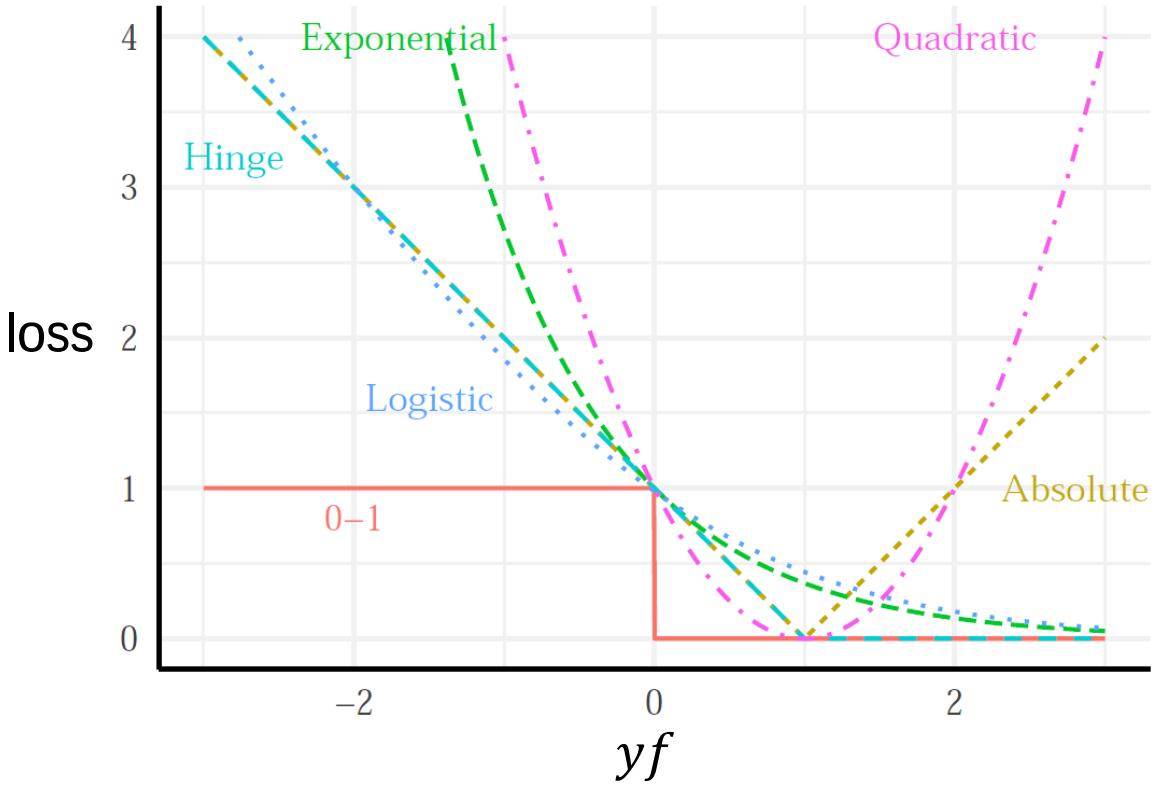


figure : Jesse H. Krijthe

Typical Classification Setting

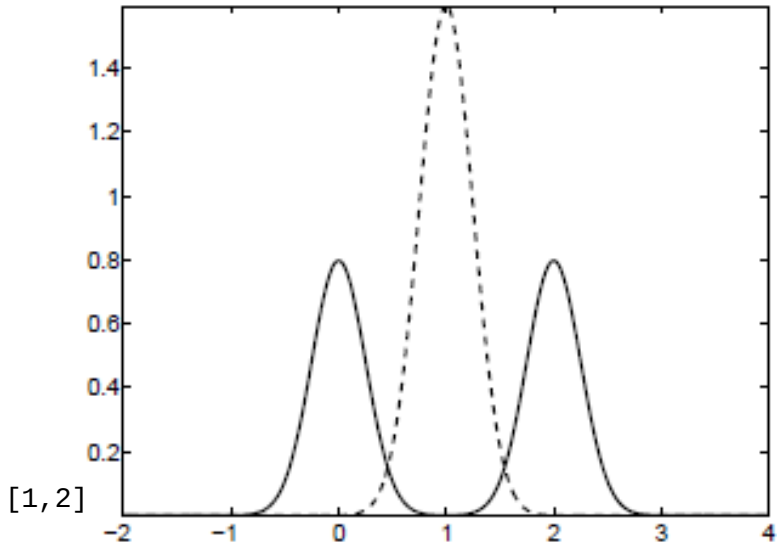
- Supervised learning comes down to determining parameters that minimize empirical surrogate loss

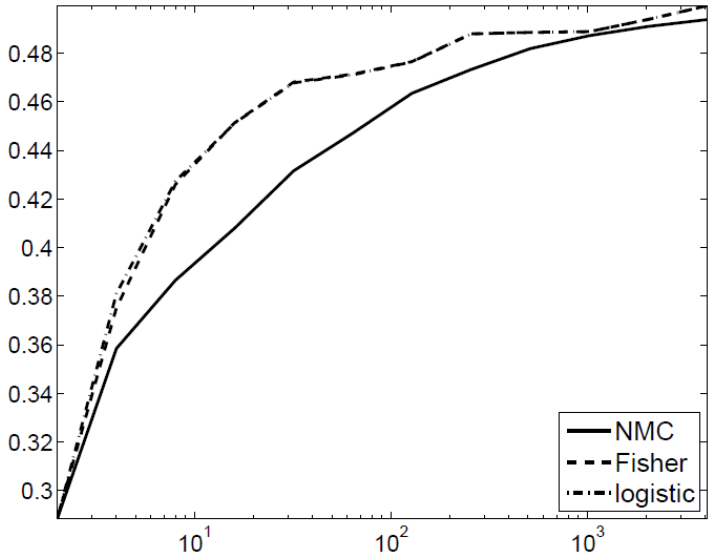
$$\theta_{\text{sup}} = \underset{\theta \in \Theta}{\text{argmin}} \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i|\theta), y_i) + R(\theta)$$

  - Possibly added is a regularizer R that penalizes overly complex solutions

More Data ⇏ Better Classifier
Continued

- Two ingredients
  - Model misspecification
  - Mismatch between risk of interest and risk optimized





Semi-Supervised Learning

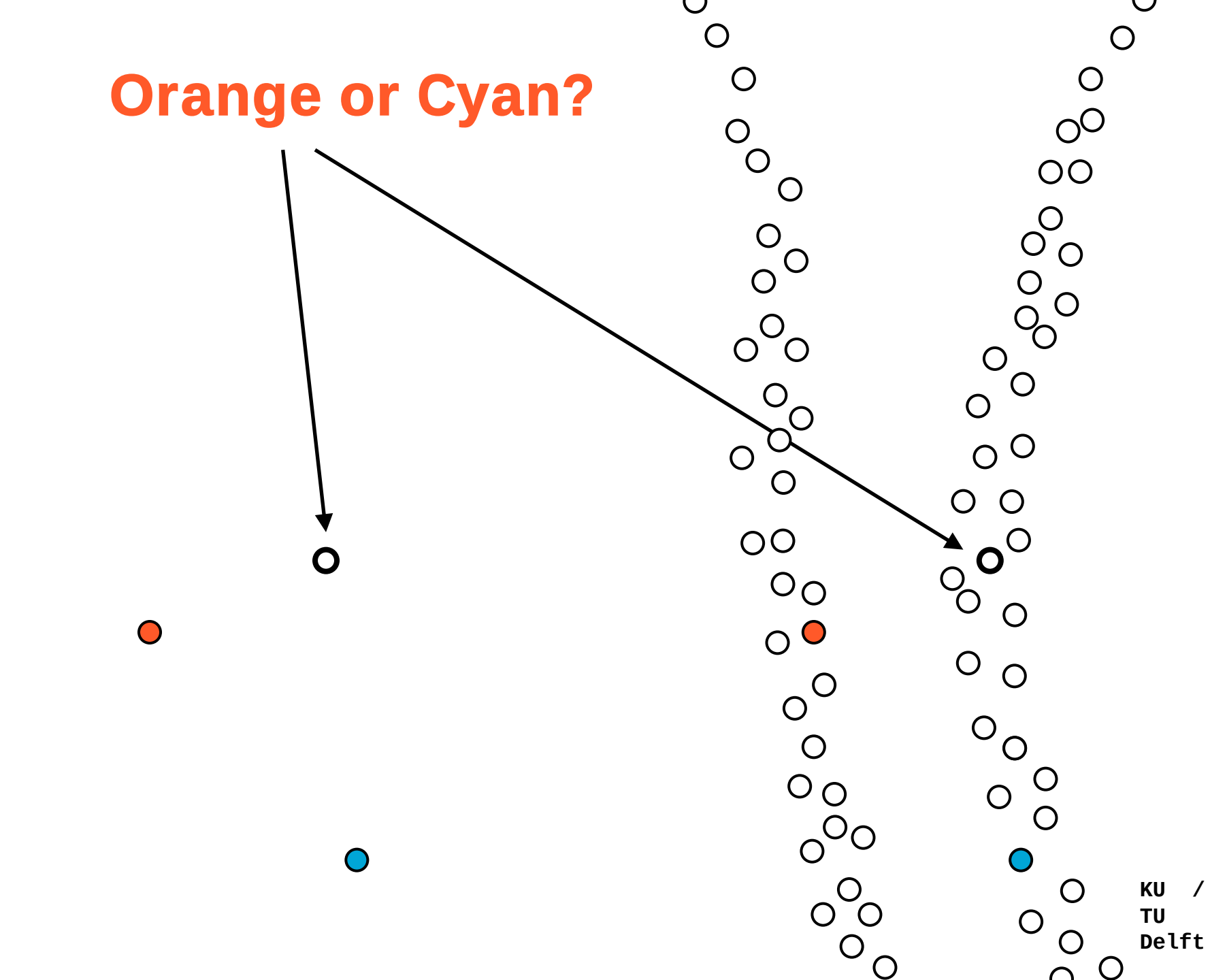
Another One of Those Basic Problems

[3,4]

Elementary Question

- How to improve a classifier when having additional unlabeled data?
  - That is, given additional sample from marginal p_X

Orange or Cyan?



KU / TU Delft

A “Technicality”

- Going to focus on transductive learning rather than semi-supervised learning
  - That is, only interest is in performance on given data [unlabeled only or both labeled and unlabeled]

[5,6]

Routes to Partial Solution

- Carefully consider Vapnik and Chervonenkis bounds
  - Focus is on error rate
- Trust choice of classifier and surrogate it optimizes
  - In a sense, don't really care about error rate

[5,6,7,8,9,10,11]

A Supervised VC Bound

- Takes on form $L_{\text{true}}(\theta) \leq L_{\text{emp}}(\theta) + C$
  - This is in terms of the 0-1 loss
  - With C a measure of complexity of hypothesis class
  - Bound is probabilistic and holds with probability $1 - \eta$

- In supervised learning, classical choice is

$$C = \sqrt{\frac{h(\log \frac{2N}{h} + 1) - \log \eta}{N}}$$

  - With h the [well-known?] VC dimension

[5,6]

A Transductive VC Bound

- Derivation of supervised bound goes through a tighter bound

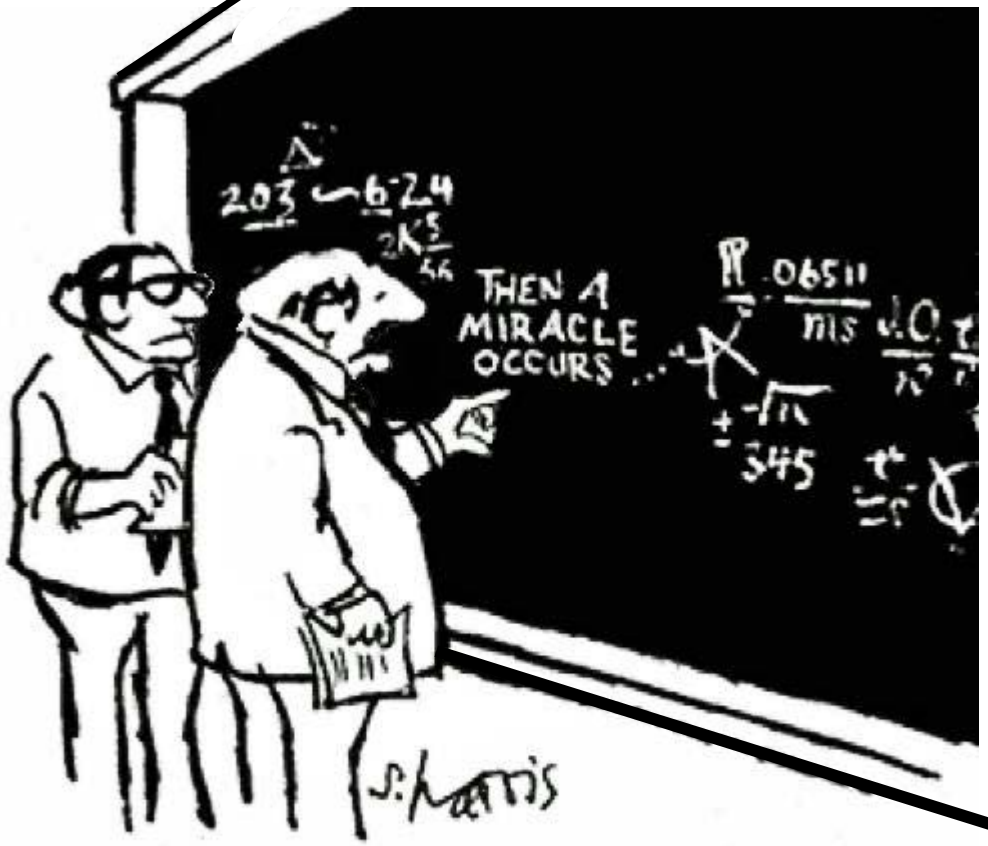
$$C = \sqrt{\frac{H_{p_X}(2N) - \log \eta}{N}} \leq \sqrt{\frac{h(\log \frac{2N}{h} + 1) - \log \eta}{N}}$$

  - Where function H_{p_X} is dependent on marginal p_X
  - This creates opportunities in transductive setting...

[5,6]

From Bound to Classifier

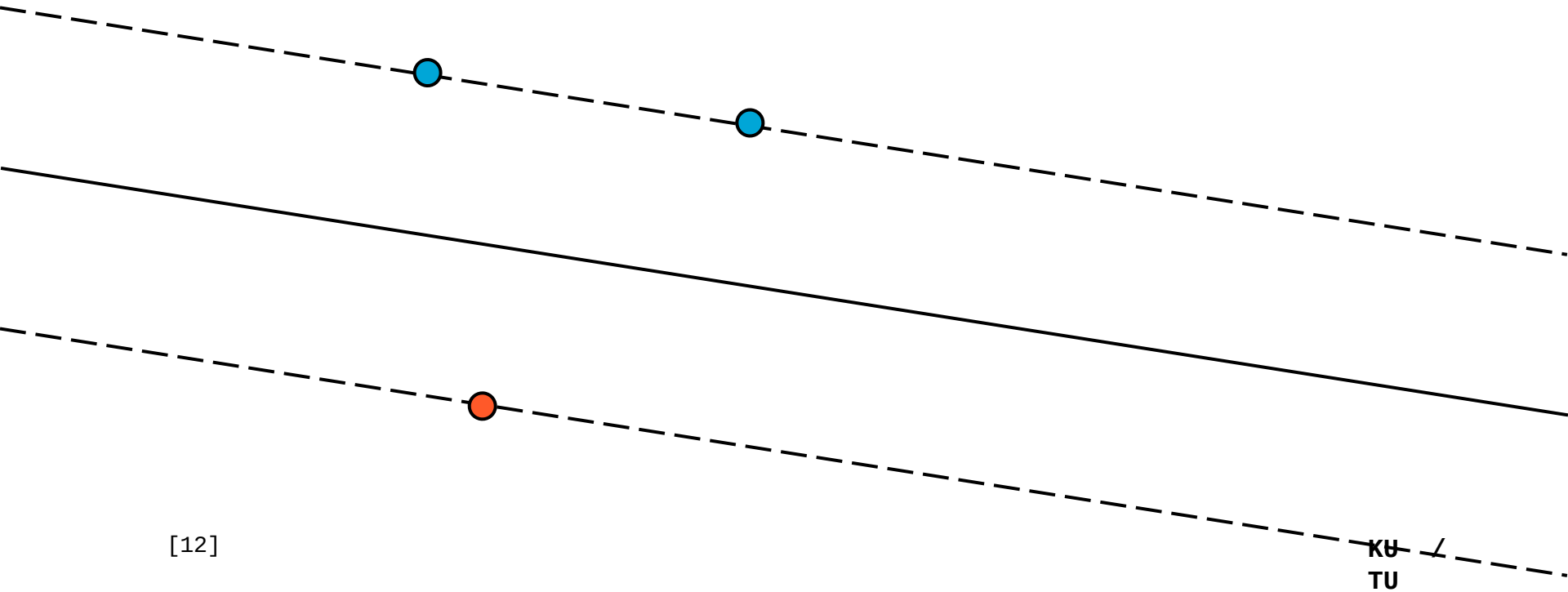
cartoon : Sidney Harris



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

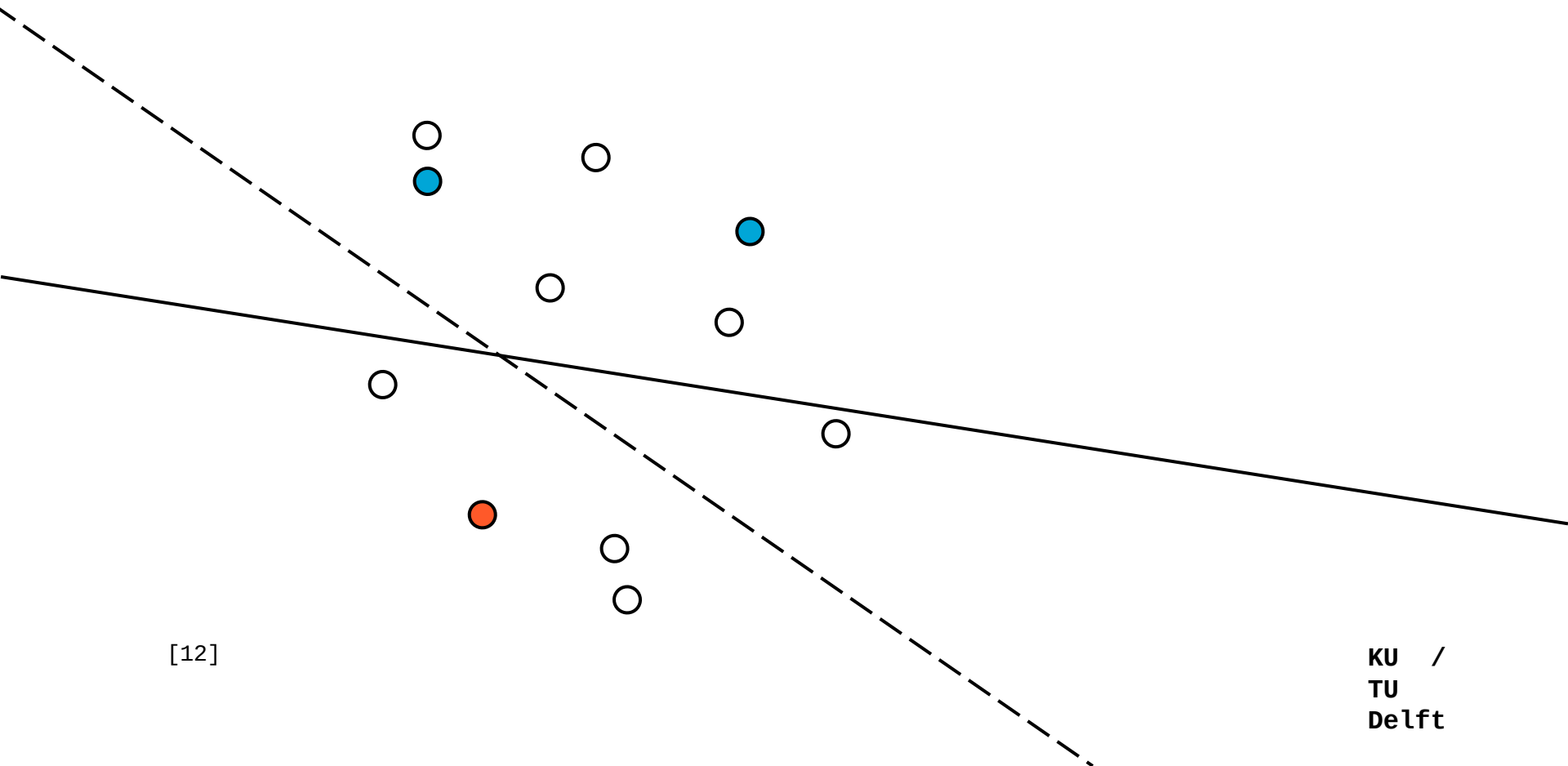
Transductive SVM

- SVM = support vector machine



Transductive SVM

- SVM = support vector machine
- Assumes that classes are linearly separable



Remarks

- Optimization of TSVM is [again] NP-hard...
- Bounds like

$$L_{\text{true}}(\theta) \leq L_{\text{emp}}(\theta) + \sqrt{\frac{H_{p_X}(2N) - \log\eta}{N}}$$

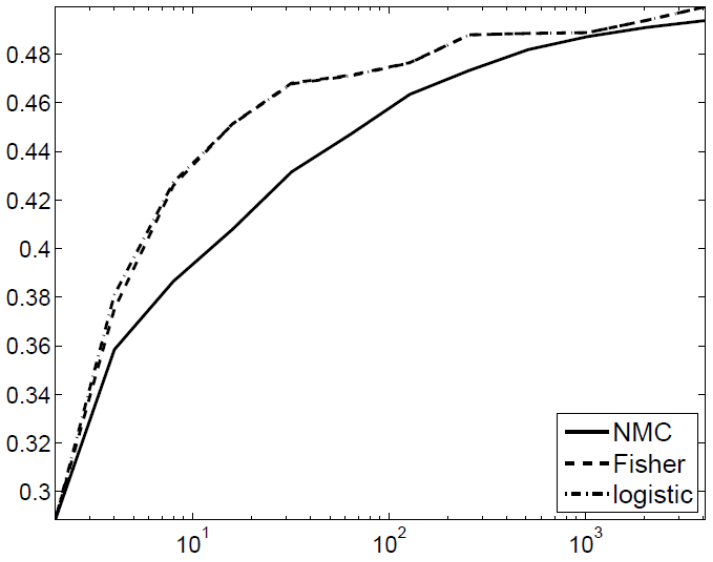
$$\leq L_{\text{emp}}(\theta) + \sqrt{\frac{h(\log\frac{2N}{h} + 1) - \log\eta}{N}}$$

do not really indicate performance improvement

  - Transductive error is not shown to become smaller...

Remarks

- But can error rate improvement be expected with transduction?
  - Well, not if we want something that is easy to compute
  - And TSVM needs separable data
  - ... this leads us to consider surrogate loss as such



- Can we get to improved supervised learner in terms of surrogate loss the classifier optimizes?

[7,8,9,10,11]

How To 1 : Contrast

- Consider [surrogate] loss on all unlabeled data, assuming one knows its labeling y_i, i.e.,

$$\frac{1}{N}\sum_{i=1}^{N}\ell(f(x_i|\theta_{\text{trn}}),y_i)$$

- Transduction not worse than supervision, then

$$\frac{1}{N}\sum_{i=1}^{N}\ell(f(x_i|\theta_{\text{trn}}),y_i)-\frac{1}{N}\sum_{i=1}^{N}\ell(f(x_i|\theta_{\text{sup}}),y_i)\leq 0$$

How To 2 : Pessimism

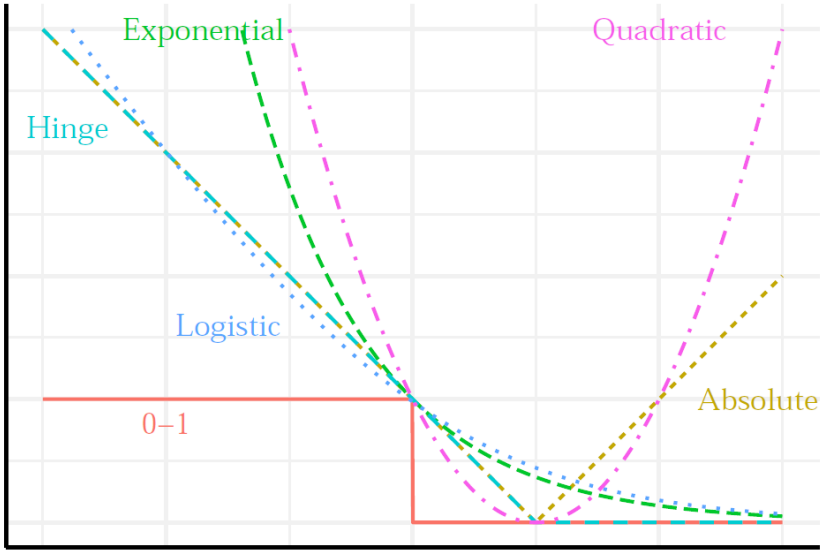
$$\frac{1}{N}\sum_{i=1}^{N}\ell(f(x_i|\theta_{\text{trn}}),y_i)-\frac{1}{N}\sum_{i=1}^{N}\ell(f(x_i|\theta_{\text{sup}}),y_i)\leq 0$$

- [Obviously] true labels are actually unknown
- Inequality for any labeling ⇒ improvement guarantee

- If possible, then following minimax gives a solution

$$\underset{\theta\in\Theta}{\text{argmin}}\ \max_{y\in\{-1,+1\}^N}\frac{1}{N}\sum_{i=1}^{N}\ell(f(x_i|\theta),y_i)-\frac{1}{N}\sum_{i=1}^{N}\ell(f(x_i|\theta_{\text{sup}}),y_i)$$

Some Results



- Using minimax construction, gets us to performance improvements for squared and absolute loss
  - Also for generative probabilistic models [log-likelihood]

- Cannot get to improvements for other losses displayed

[9,10,11,13]

General Status of Transduction

- Stronger bounds for 0-1 loss
  - True error can be more tightly related to empirical error
  - Does not guarantee improvements over supervised classifier
- Focusing on surrogate losses
  - Some classifiers can provide guaranteed improvements
  - Other classifiers can, provably, not be improved
- For SVMs in particular
  - Cannot guarantee improvement in terms of surrogate loss
  - TSVM is NP-hard and “needs” separable data assumption

Some Related Problems

Quick? I Probably Don’t Really Have Time for This…

[14,15,16,17]

Similar Issues

- Many settings where similar issues concerning computation and surrogate losses play a role
  - Active learning : can we label data in a wiser way than by just drawing independent realizations?
  - Transfer learning / domain adaptation : what if the density p_{XY} is different between training and test phase?

[14,15,16,17]

Speculations

How Quantum Computing Affects Machine Learning?

KU / TU Delft

What If We Could

- Find the actual minimizer for TSVMs?
  - Does that solve the semi-supervised learning problem?
- Empirically investigate global solutions to nonconvex losses?
  - Interesting to see how surrogate solutions really differ
- Globally optimize neural networks?
  - Finally we see that they indeed heavily overtrain...!
- Compare as many models as we would like...?

- Would we discover the benefits of non-optimality?

Qs?

...and Maybe Some As

KU / TU Delft

A Mildly Biased Reference List

- [1] Loog, Duin, "The dipping phenomenon," S+SSPR, 2012
- [2] Loog, Krijthe, Jensen, "On measuring and quantifying performance," Handbook of PRCV, World Scientific
- [3] Chapelle, Schölkopf, Zien, "Semi-Supervised Learning," MIT press, 2006
- [4] Zhu, "Semi-supervised learning literature survey," University of Wisconsin, TR 1530, 2008
- [5] Vapnik, "Statistical Learning Theory," John Wiley & Sons, 1998
- [6] Vapnik, "Transductive Inference and Semi-Supervised Learning," Chapter 24 in [3]
- [7] Loog, "Constrained parameter estimation for semi-supervised learning", ECML, 2010
- [8] Loog, "Semi-supervised linear discriminant analysis through moment-constraint parameter estimation," PRL, 2014
- [9] Loog, "Contrastive Pessimistic Likelihood Estimation for Semi-Supervised Classification", IEEE TPAMI, 2016
- [10] Krijthe, Loog, "Robust semi-supervised least squares classification by implicit constraints," PR, 2017
- [11] Krijthe, Loog, "Projected Estimators for Robust Semi-supervised Classification," ML, 2017
- [12] Joachims, "Transductive inference for text classification using support vector machines," ICML 1999
- [13] Krijthe, Loog, "The Pessimistic Limits of Margin-based Losses in Semi-supervised Learning," arXiv, 2016
- [14] Settles, “Active Learning Literature Survey”, TR 1648, University of Wisconsin–Madison, 2010
- [15] Loog, Yang, “An Empirical Investigation into the Inconsistency of Sequential Active Learning,” ICPR 2016
- [16] Quionero-Candela, Sugiyama, Schwaighofer, Lawrence, “Dataset shift in machine learning,” The MIT Press, 2009
- [17] Kouw, Loog, "Target contrastive pessimistic risk for robust domain adaptation," arXiv, 2017
- [18] Harris, "What's So Funny About Science?" William Kaufmann, 1977