Behavior of Distance-Based Methods
in a Context of Class-Imbalance or High-Dimensionality

Rémi Emonet

Université Jean-Monnet, Laboratoire Hubert Curien, Saint-Étienne

Talk at Tahiti (Bréhat), 2019-06-27

$whoami

Overview

- <u>Introduction</u>
- High-dimensional problems
  - The curse of dimensionality
  - Ockham's Razor
  - Notions of Simplicity
- High-dimensionality and Neighborhood
- Imbalanced classification problems
  - The Problem (and performance measures)
  - Reweight, resampling, etc
  - Correcting k-NN (γ-***NN***)
  - Focusing on the F-Measure optimization (Élisa)
- Discussion

Overview

- Introduction
- High-dimensional problems
  - The curse of dimensionality
  - Ockham's Razor
  - Notions of Simplicity
- High-dimensionality and Neighborhood
- Imbalanced classification problems
  - The Problem (and performance measures)
  - Reweight, resampling, etc
  - Correcting k-NN (γ-***NN***)
  - Focusing on the F-Measure optimization (Élisa)
- Discussion

The Curse of Dimensionality

High-dimensionality ~~is~~ can be a mess.

What is this Curse Anyway?

- Some definition:

  Various phenomena that arise
  when analyzing and organizing data
  in high-dimensional spaces.

- Term coined by Richard E. Bellman
  - 1920 – 1984
  - dynamic programming
  - differential equations
  - shortest path
- What is (not) the cause?
  - not an intrinsic property of the data
  - depends on the representation
  - depends on how data is analyzed

Combinatorial Explosion

- Suppose
  - you have d entities
  - each can be in 2 states
- Then
  - there are 2^d combinations to consider/test/evaluate
- Happens when considering
  - all possible subsets of a set (2^d)
  - all permutations of a list $(d!)$
  - all affectations of entities to labels (k^d, with k labels)

{a}

```
{   }
{ a }
```

{a,b}

```
{       }
{     b }
{ a     }
{ a , b }
```

{a,b,c}

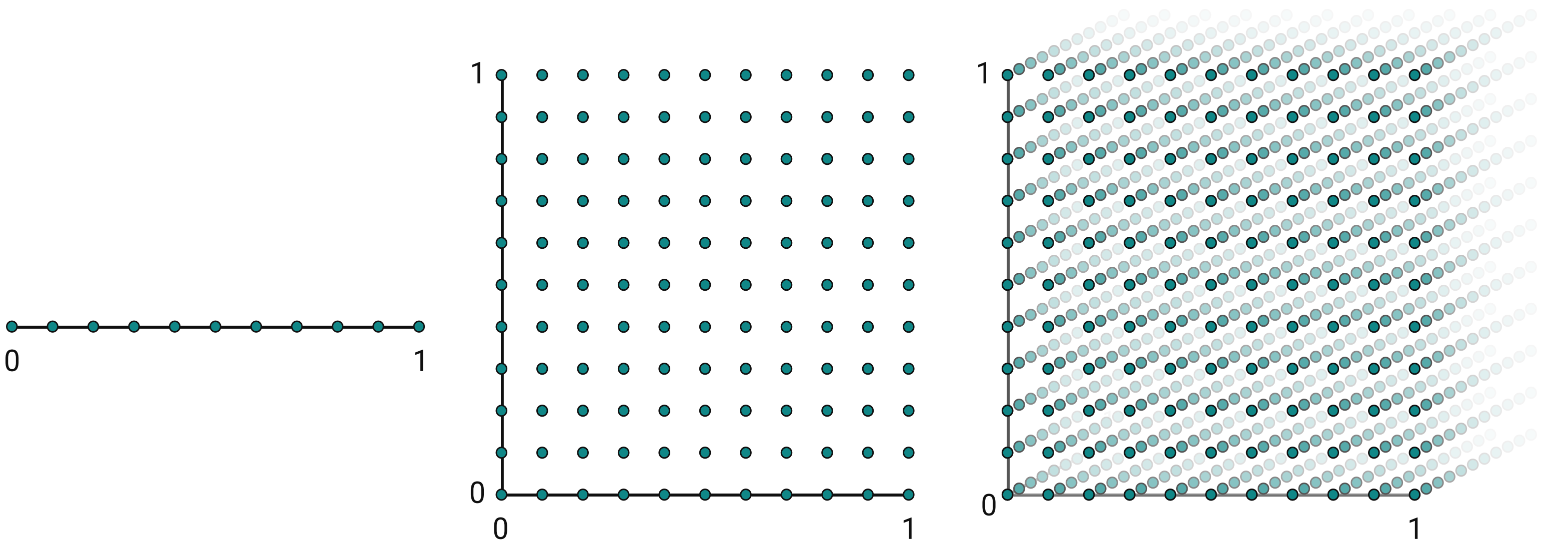
```
{           }
{         c }
{     b     }
{     b , c }
{ a         }
{ a ,     c }
{ a , b     }
{ a , b , c }
```

{a,b,c,d}

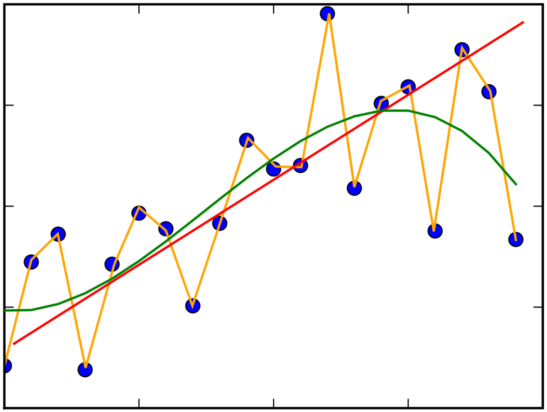
```
{               }
{             d }
{         c     }
{         c , d }
{     b         }
{     b ,     d }
{     b , c     }
{     b , c , d }
{ a             }
{ a ,         d }
```

Regular Space Coverage

- Analogous to combinatorial explosion, in continuous spaces
- Happens when considering
  - histograms
  - density estimation
  - anomaly detection
  - ...



In Modeling and Learning



- The world is complicated
  - state with a huge number of variables (dimensions)
  - possibly noisy observations
  - e.g. a 1M-pixel image has 3 million dimensions
- Learning would need observations for each state
  - it would require too many examples
  - need for an “interpolation” procedure, to avoid overfitting
- Hughes phenomenon, 1968 paper (which is wrong, it seems)

  given a (small) number of training samples,

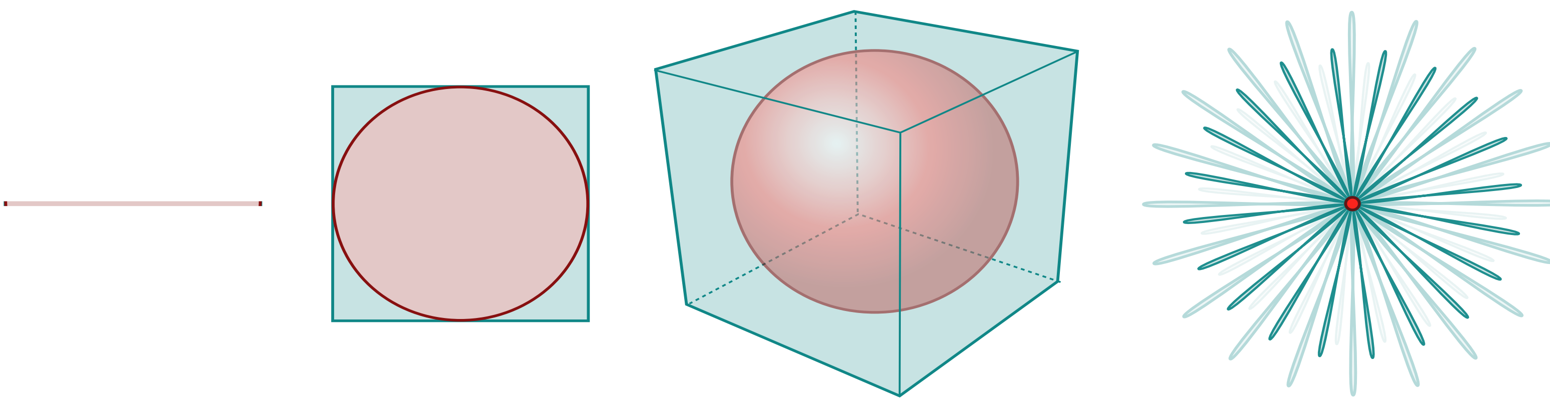
  additional feature measurements

  may reduce the performance of a statistical classifier

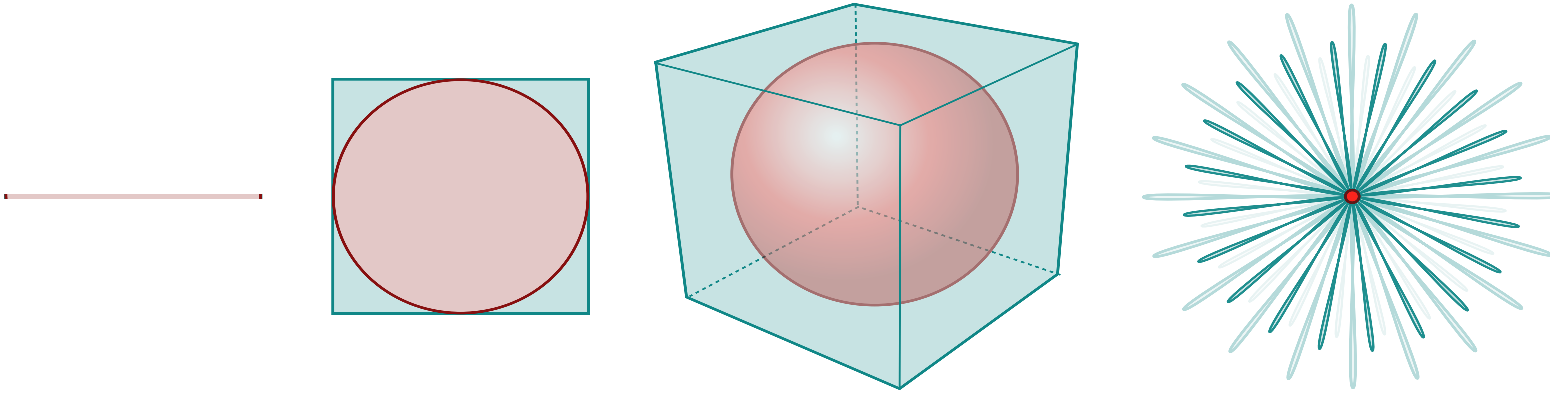
A Focus on Distances/Volumes

- Considering a d dimensional space
- About volumes
  - volume of the cube: $C_d(r) = (2r)^d$
  - volume of a sphere with radius r: $S_d(r) = \dfrac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} r^d$

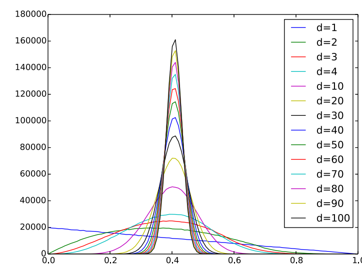
    (Γ is the continuous generalization of the factorial)
  - ratio: $\dfrac{S_d(r)}{C_d(r)} \rightarrow 0$ (linked to space coverage)



A Focus on Distances/Volumes (cont'd)



- About distances
  - average (euclidean) distance between two random points?
  - everything becomes almost **as** “far”
- Happens when considering
  - radial distributions (multivariate normal, etc)
  - k-nearest neighbors (hubness problem)
  - other distance-based algorithms



The Curse of Dimensionality

Many things get degenerated with high dimensions
Problem of: approach + data representation
We have to hope that there is no curse

Overview

- Introduction
- High-dimensional problems
  - The curse of dimensionality
  - Ockham's Razor
  - Notions of Simplicity
- High-dimensionality and Neighborhood
- Imbalanced classification problems
  - The Problem (and performance measures)
  - Reweight, resampling, etc
  - Correcting k-NN (γ-***NN***)
  - Focusing on the F-Measure optimization (Élisa)
- Discussion

Ockham's Razor

Shave unnecessary assumptions.

Ockham's Razor

- Term from 1852, in reference to Ockham (XIVth)
- *lex parsimoniae*, law of parsimony
- *Prefer the simplest hypothesis that fits the data.*
- Formulations by Ockham, but also earlier and later
- More a concept than a rule
  - simplicity
  - parsimony
  - elegance
  - shortness of explanation
  - shortness of program (Kolmogorov complexity)
  - falsifiability (sciencific method)
- According to Jürgen Schmidhuber, *the appropriate mathematical theory of Occam's razor already exists, namely, Solomonoff's theory of optimal inductive inference.*

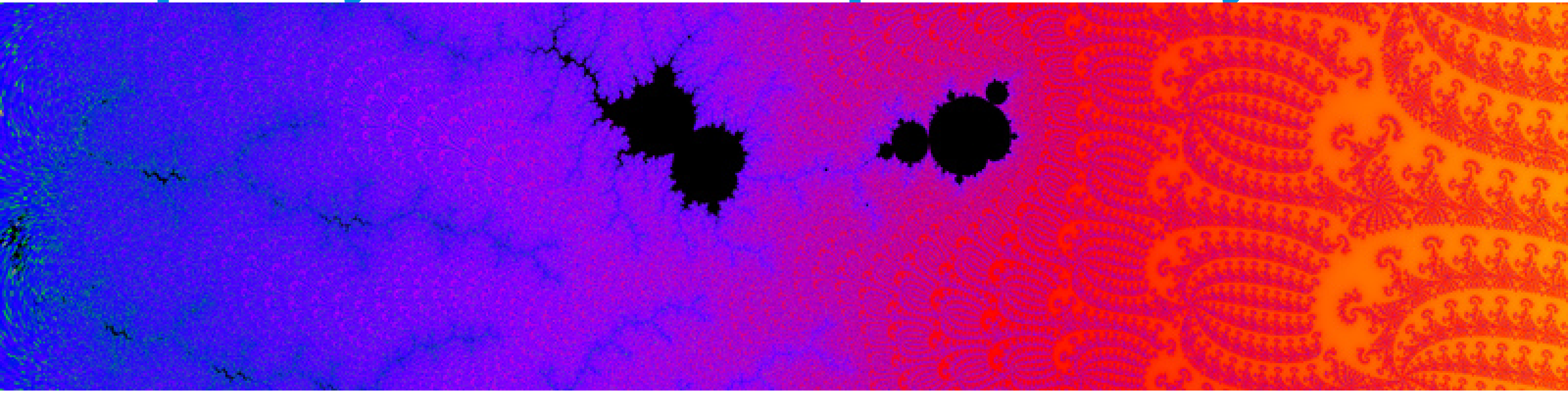
Overview

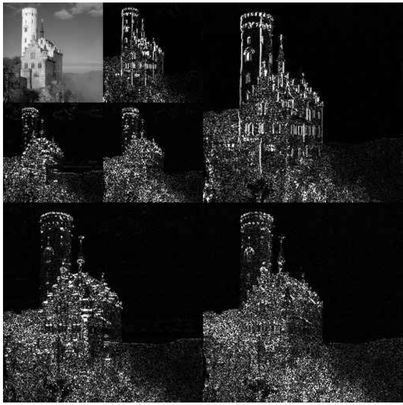
- Introduction
- High-dimensional problems
  - The curse of dimensionality
  - Ockham's Razor
  - Notions of Simplicity
- High-dimensionality and Neighborhood
- Imbalanced classification problems
  - The Problem (and performance measures)
  - Reweight, resampling, etc
  - Correcting k-NN (γ-***NN***)
  - Focusing on the F-Measure optimization (Élisa)
- Discussion

Simplicity of Data: subspaces

- Data might be high-dimensional, but we have hope
  - that there is a organization or regularity in the high-dimensionality
  - that we can guess it
  - or, that we can learn/find it
- Approaches: dimensionality reduction, manifold learning
  - PCA, kPCA, *PCA, SOM, Isomap, GPLVM, LLE, NMF, …

Simplicity of Data: compressibility





- Idea
  - data can be high dimensional but compressible
  - i.e., there exist a compact representation
- Program that generates the data (Kolmogorov complexity)
- Sparse representations
  - wavelets (jpeg), fourier transform
  - sparse coding, representation learning
- Minimum description length
  - size of the “code” + size of the encoded data

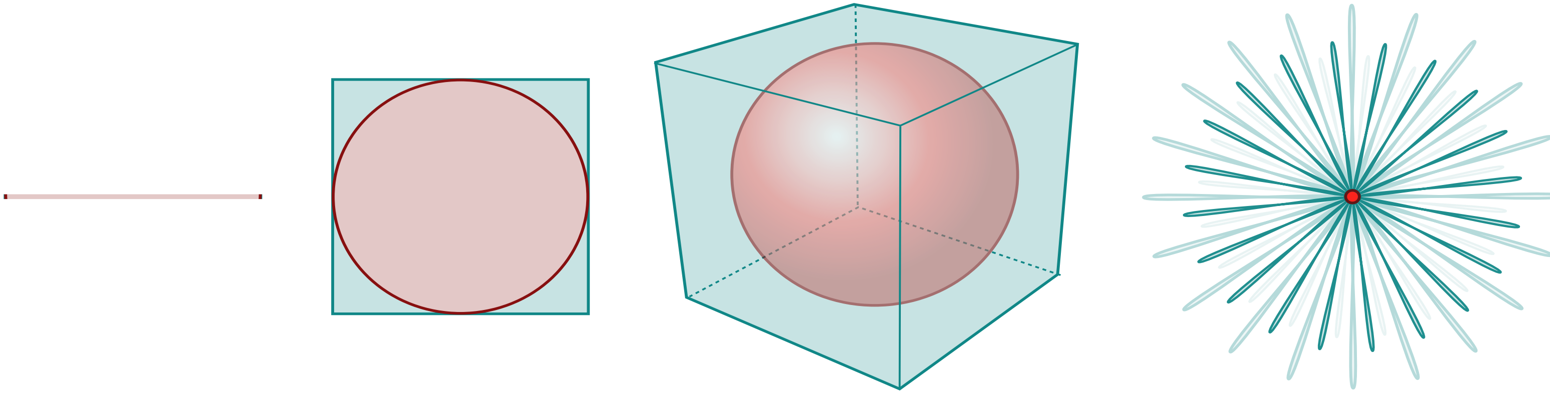
Simplicity of Models: information criteria

- Used to select a model
- Penalizes by the number k of *free parameters*
  - AIC (Aikake Information Criterion)
    - penalizes the Negative-Log-Likelihood by k
  - BIC (Bayesian IC)
    - penalizes the NLL by $k \log(n)$ (for n observations)
  - BPIC (Bayesian Predictive IC)
  - DIC (Deviance IC)
  - FIC (Focused IC)
  - Hannan-Quinn IC
  - TIC (Takeuchi IC)
- Sparsity of the parameter vector ($l0$ norm)
  - penalizes the number of non-zero parameters

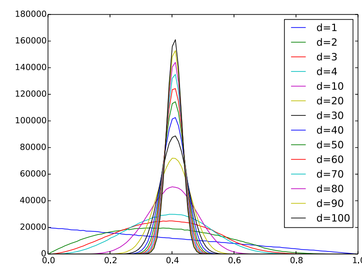
Overview

- Introduction
- High-dimensional problems
    - The curse of dimensionality
    - Ockham's Razor
    - Notions of Simplicity
- High-dimensionality and Neighborhood
- Imbalanced classification problems
    - The Problem (and performance measures)
    - Reweight, resampling, etc
    - Correcting k-NN (γ-***NN***)
    - Focusing on the F-Measure optimization (Élisa)
- Discussion

A Focus on Distances/Volumes (cont'd)

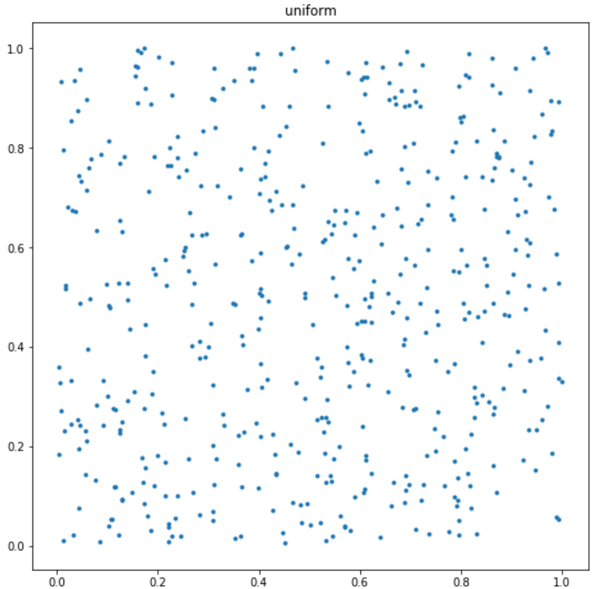


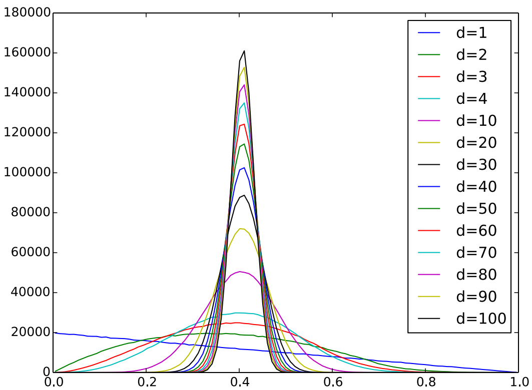
- About distances
  - average (euclidean) distance between two random points?
  - everything becomes almost **as** “far”
- Happens when considering
  - radial distributions (multivariate normal, etc)
  - k-nearest neighbors (hubness problem)
  - other distance-based algorithms



Distance Contraction

- Experiment
  - sampling uniformly random points in the unit cube
  - looking at the distribution of inter-point distances
  - variance decreases with dimensionality

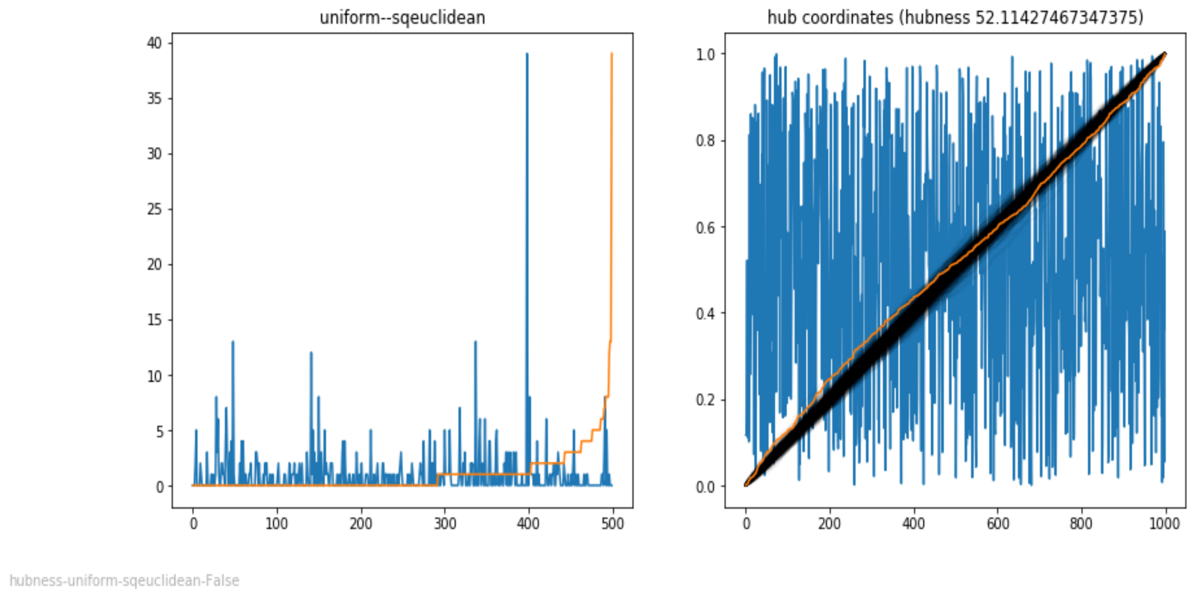




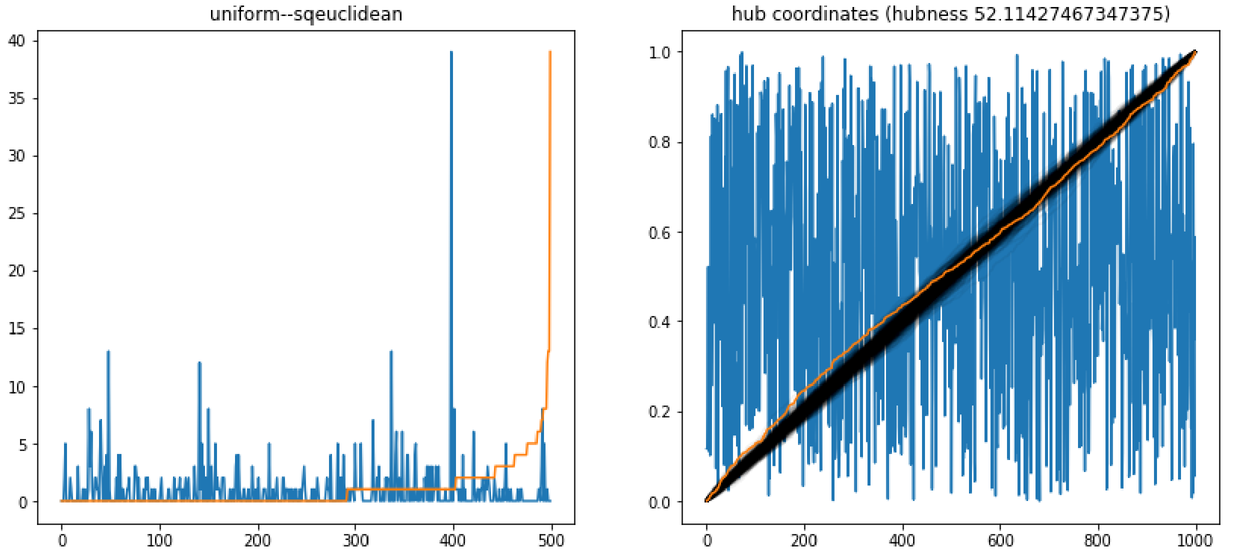
- Question: is it a problem? *maybe not if the ranking is right*

Hubness Problem

- Experiment
  - sampling uniformly random points in the unit cube
  - computing how often each point
    is in the nearest neighbor of another point
  - Hubness as skewness: $hubness = \frac{\mathbb{E}[(N - \mu_N)^3]}{\sigma_N^3}$



Hubness: why are some points so special?

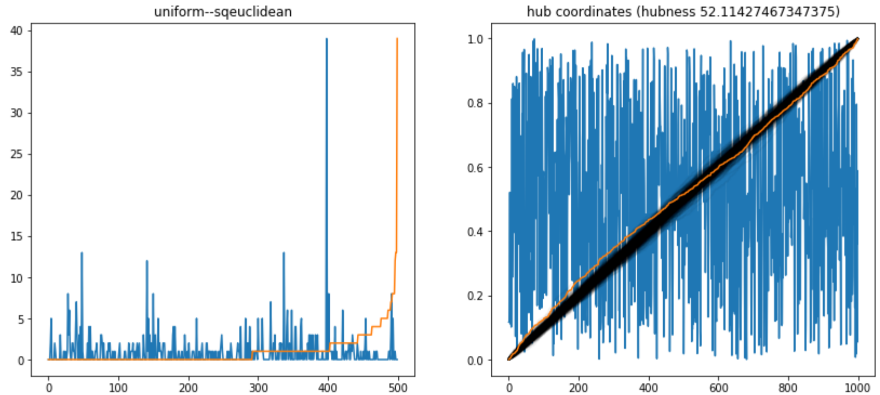


hubness-uniform-sqeuclidean-False

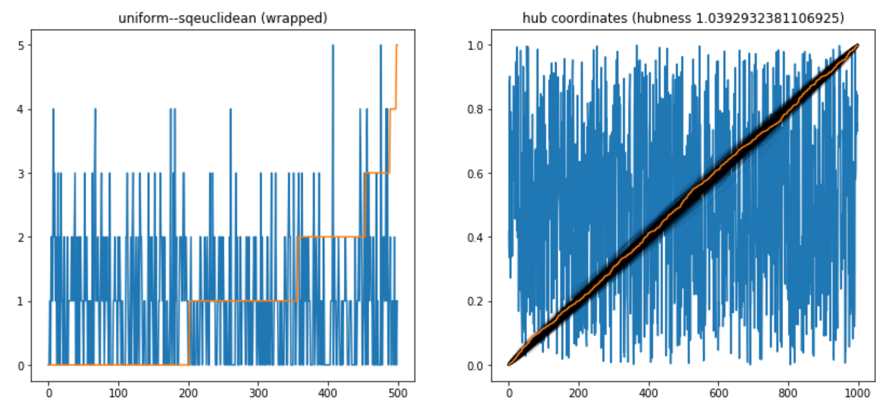
- Where are these points?
- The border theory...
- ... so it is distribution-dependant

Hubness: testing the border theory

- Wrapping the points (hyper-torus)



hubness-uniform-sqeuclidean-False



hubness-uniform-sqeuclidean-True

Hubness: what is a border?

Overview

- Introduction
- High-dimensional problems
  - The curse of dimensionality
  - Ockham's Razor
  - Notions of Simplicity
- High-dimensionality and Neighborhood
- Imbalanced classification problems
  - The Problem (and performance measures)
  - Reweight, resampling, etc
  - Correcting k-NN (γ-***NN***)
  - Focusing on the F-Measure optimization (Élisa)
- Discussion

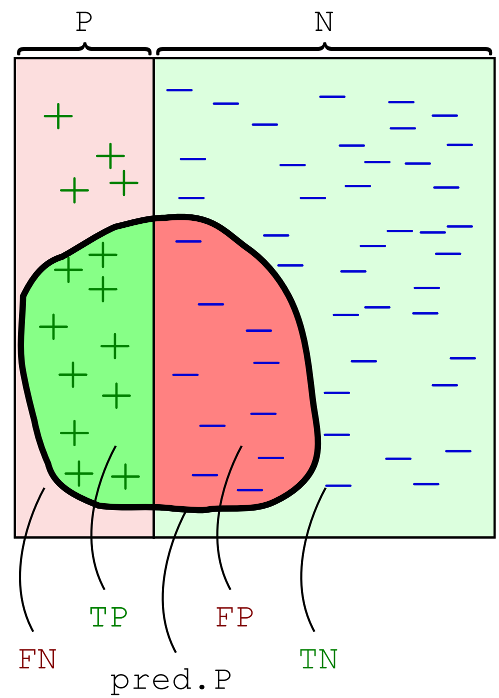
Imbalanced Problems: Examples

- Anomaly detection
  - unsafe situations in videos
  - defect detection in images
  - abnormal heart beat detection in ECG
- Fraud detection
  - fraudulent checks
  - credit card fraud (physical, online)
  - financial fraud (French DGFIP)

Imbalanced Classification Problems

- Binary classification
  - $+$ positive class: minority class, anomaly, rare event, …
  - $-$ negative class: majority class, normality, typical event, …
- Confusion matrix (of a model vs a ground truth)
  - TP: true positive
  - FP: false positive
  - TN: true negative
  - FN: false negative
- Some measures
  - Precision: $prec = \dfrac{TP}{TP + FP}$
  - Recall: $rec = \dfrac{TP}{P} = \dfrac{TP}{TP + FN}$
  - F_β-measure: $F_\beta = (1 + \beta^2)\dfrac{prec \cdot rec}{\beta^2 \cdot prec + rec}$

  *(higher is better)



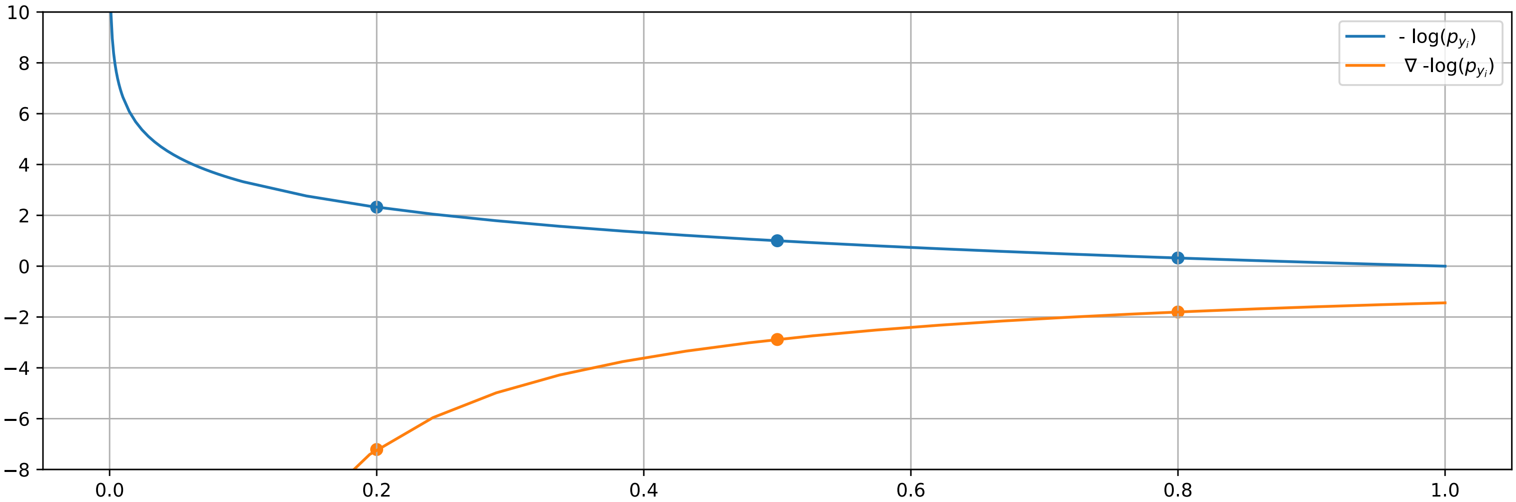
F-measure vs Accuracy ?

$$F_\beta = (1+\beta^2)\frac{prec \cdot rec}{\beta^2 \cdot prec + rec} = \frac{(1+\beta^2)\cdot(P-FN)}{1+\beta^2 P - FN + FP}$$

$$accuracy = \frac{TP+TN}{P+N} = 1 - \frac{FN+FP}{P+N}$$

- Accuracy inadequacy (e.g. $N = 10000, P = 100$)
  - lazy "all$-$" classifier $(TP = 0, TN = N, FP = 0, FN = P)$
  - $accuracy = \frac{0+N}{P+N} = \frac{10000}{10100} = 99\%$
  - $F_\beta = \frac{(1+\beta^2)(P-P)}{1+\beta^2 P - P + 0} = 0$
- F_β-measure challenges
  - discrete (like the accuracy)
  - non-convex (even with continuous surrogates)
  - **non-separable**, i.e. $F_\beta \neq \sum_{(x_i,y_i)\in S} \dots$

Ok, but I'm doing gradient descent, so ...



- Gradient: $0.2 \Rightarrow -7.21, \quad 0.5 \Rightarrow -2.89, \quad 0.8 \Rightarrow -1.80, \quad 1 \Rightarrow -1.44$
- Example, gradient intensity is the same for:
  - $10+$ wrongly classified with an output proba. of 0.2
  - $40-$ correctly classified with an output proba 0.8
  - i.e., lazily predicting systematically 0.2 (for $+$) yields a "stable" solution with $10+$ vs $40-$

Overview

- Introduction
- High-dimensional problems
    - The curse of dimensionality
    - Ockham's Razor
    - Notions of Simplicity
- High-dimensionality and Neighborhood
- Imbalanced classification problems
    - The Problem (and performance measures)
    - Reweight, resampling, etc
    - Correcting k-NN (γ-***NN***)
    - Focusing on the F-Measure optimization (Élisa)
- Discussion

Counteracting Imbalance

- Undersampling the majority class $-$
- Oversampling class $+$
- Generating fake $+$
- Using a weighted-classifiers learner

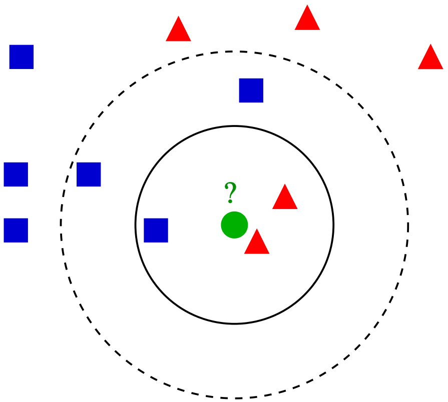
Overview

- Introduction
- High-dimensional problems
  - The curse of dimensionality
  - Ockham's Razor
  - Notions of Simplicity
- High-dimensionality and Neighborhood
- Imbalanced classification problems
  - The Problem (and performance measures)
  - Reweight, resampling, etc
  - Correcting k-NN (γ-***NN***)
  - Focusing on the F-Measure optimization (Élisa)
- Discussion

A Corrected Nearest Neighbor Algorithm Maximizing the F-Measure from Imbalanced Data

- **Rémi Viola**, Rémi Emonet , Amaury Habrard, **Guillaume Metzler**, Sébastien Riou, Marc Sebban
- ???

k-NN: k Nearest Neighbor Classification



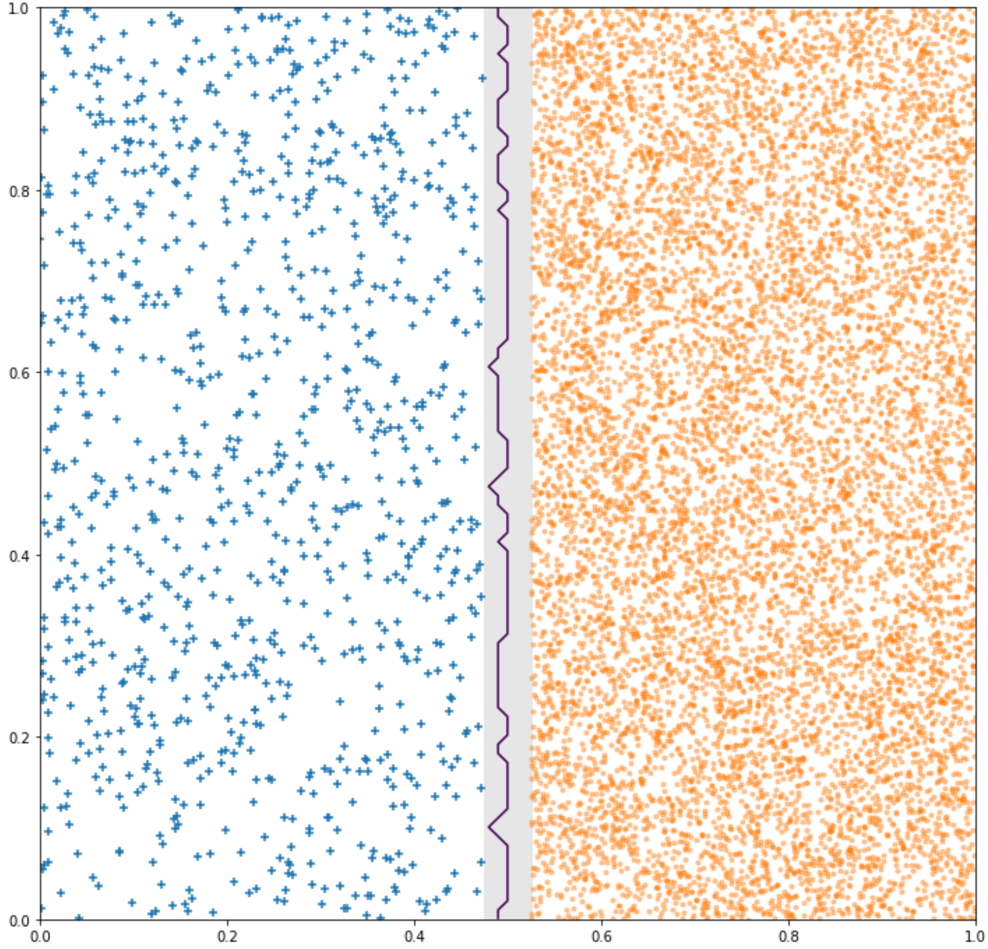
- k-NN
  - to classify a new point
  - find the closest k points (in the training section)
  - use a voting scheme to affect a class
  - efficient algorithms (K-D Tree, Ball Tree)
- Does k-NN still matter?
  - non-linear by design (with similarity to RBF-kernel SVM)
  - no learning, easy to patch a model (add/remove points)
  - Limits of k-NN for imbalanced data?

Limits of k-NN for imbalanced data?

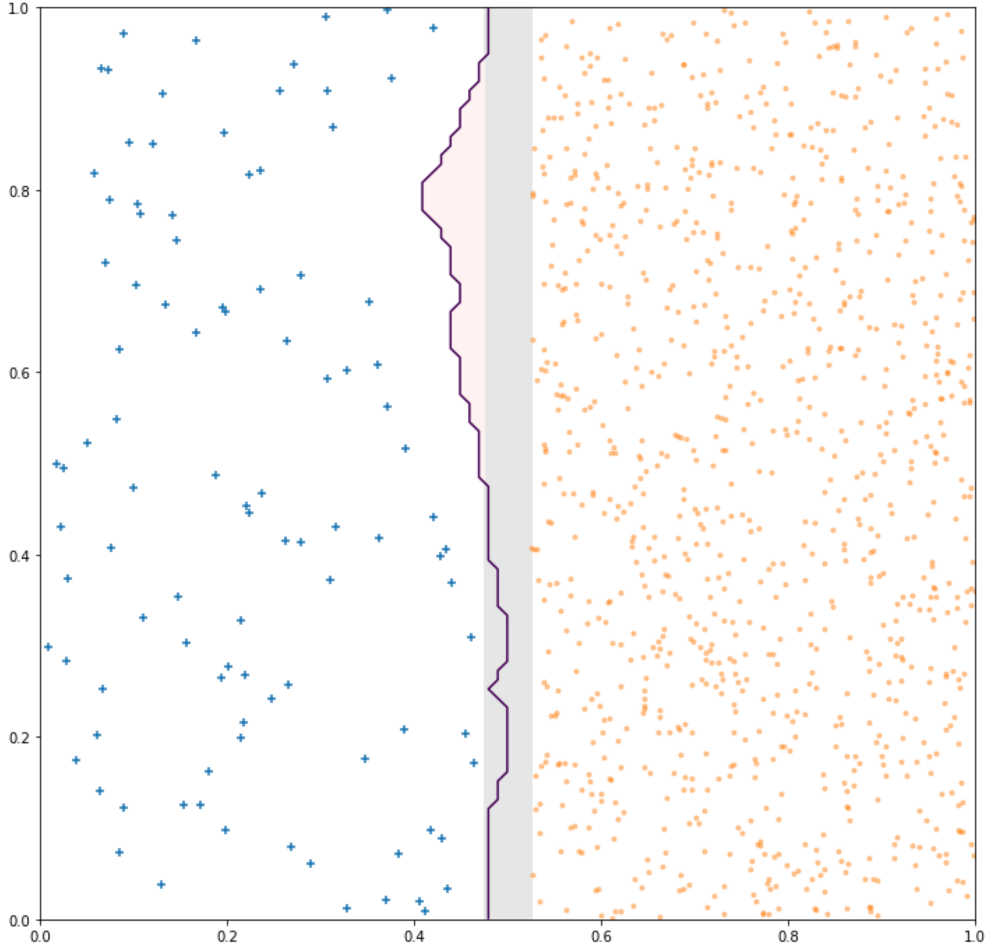
1. k-NN behavior in uncertain areas
   - i.e., for some feature vector, the class can be $+$ or $-$
   - i.e., the Bayes Risk is non zero
   - ✔ not so bad (respects imbalance)

2. k-NN behavior around boundaries
   - i.e., what happens if classes are separate but imbalanced
   - ✖ sampling effects cause problems

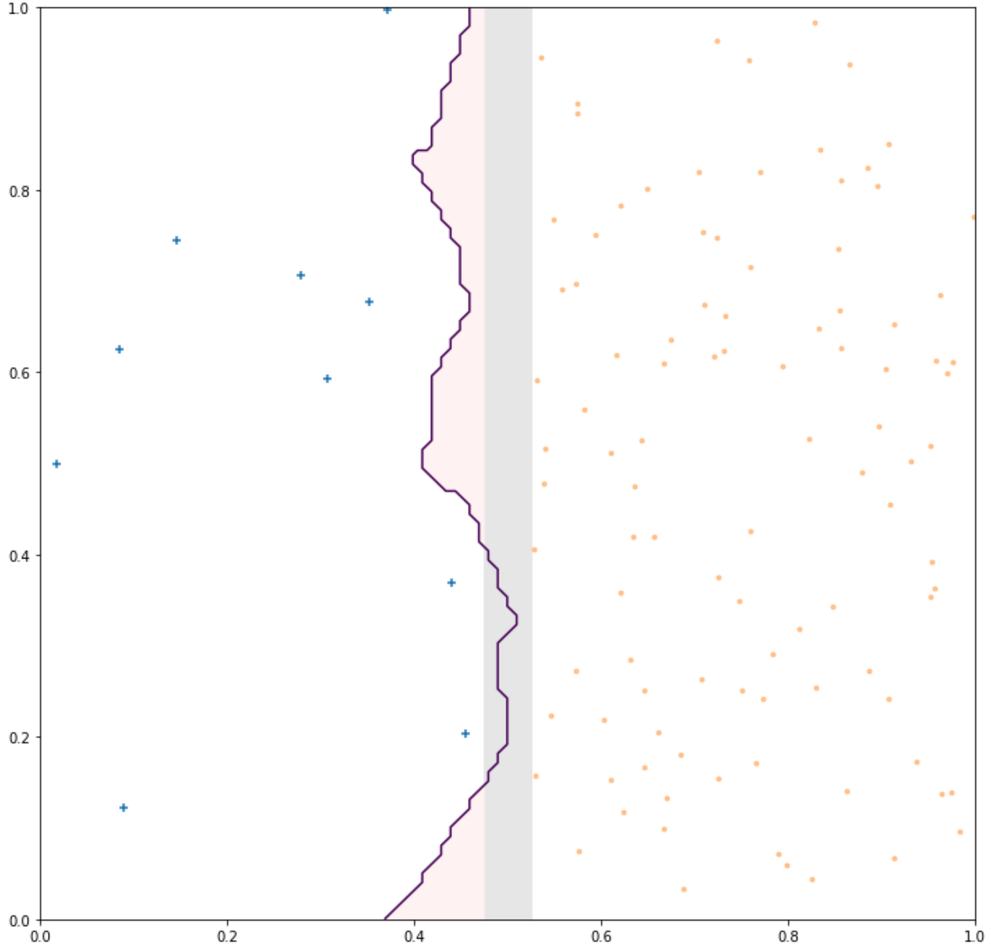
k-NN at a boundary (1000 +)



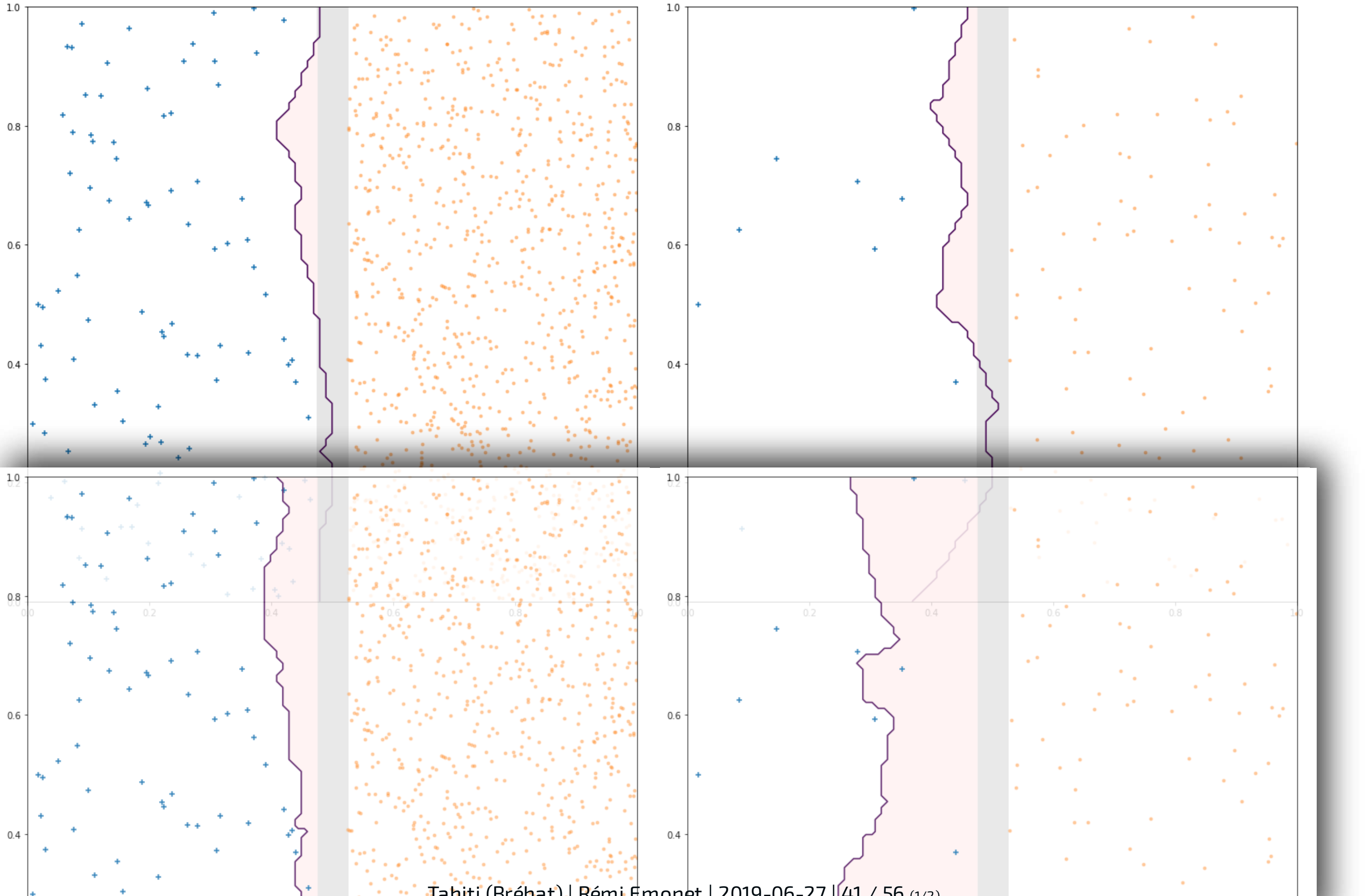
k-NN at a boundary (100 $+$)



k-NN at a boundary (10 $+$)



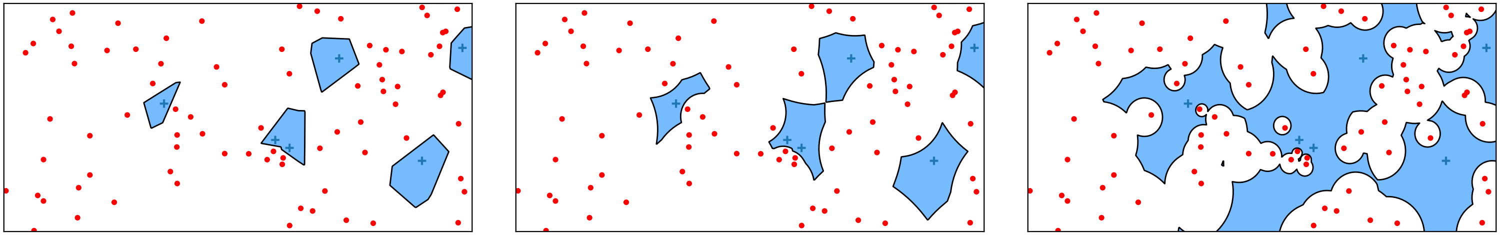
k-NN: increasing k?



A Corrected Nearest Neighbor Algorithm Maximizing the F-Measure from Imbalanced Data

- **Rémi Viola**, Rémi Emonet , Amaury Habrard, **Guillaume Metzler**, Sébastien Riou, Marc Sebban
- ???

γ-NN Idea: push the decision boundary



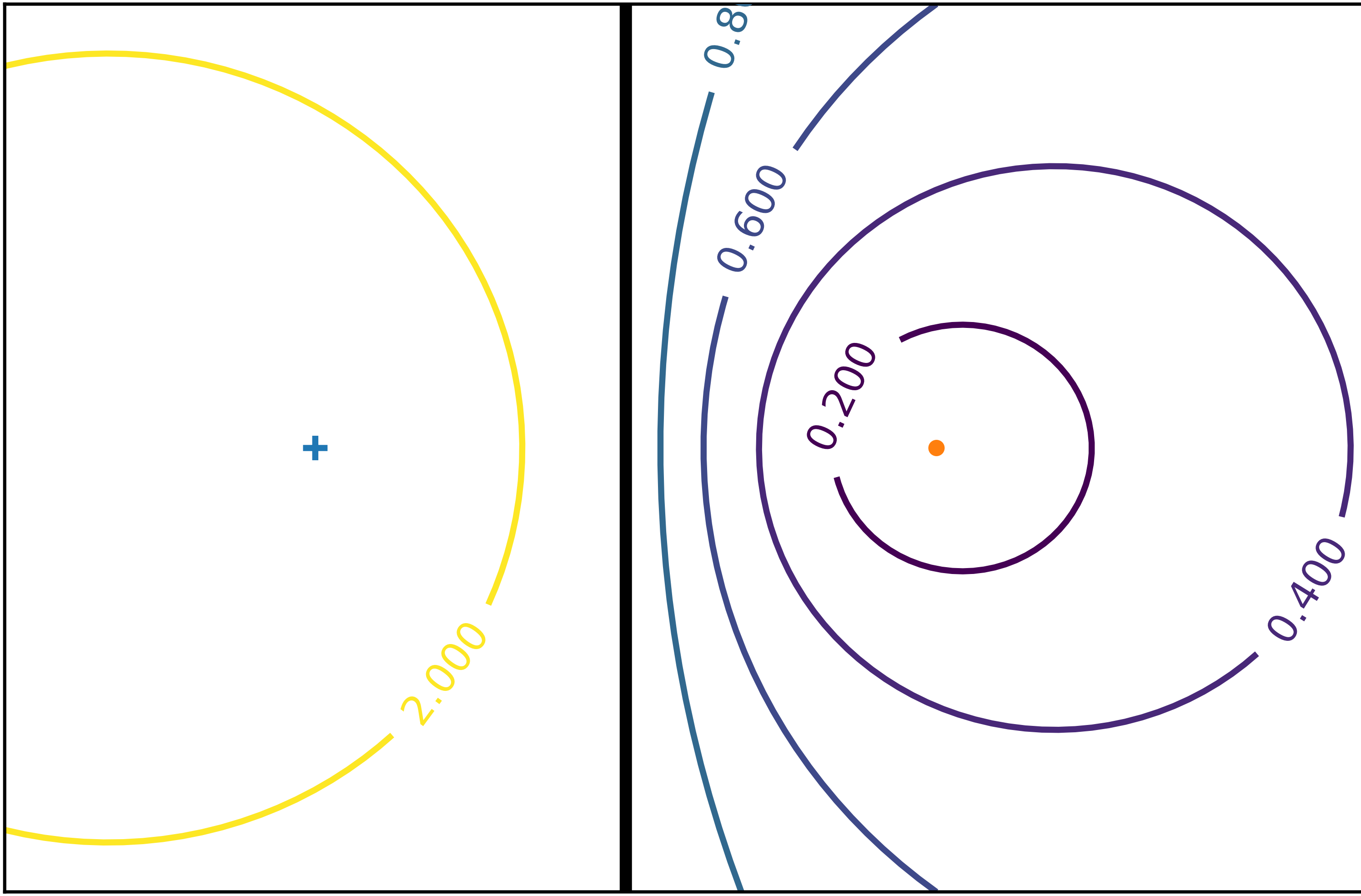
- Goal: correct for problems due to sampling with imbalance
- Genesis: GAN to generate "$+$" around existing ones

  $\Rightarrow$ unstable, failing, complex
- Approach
  - artificially make $+$ closer to new points
  - how? by using a different distance for $+$ and $-$
  - the base distance to $+$ gets multiplied by a parameter γ

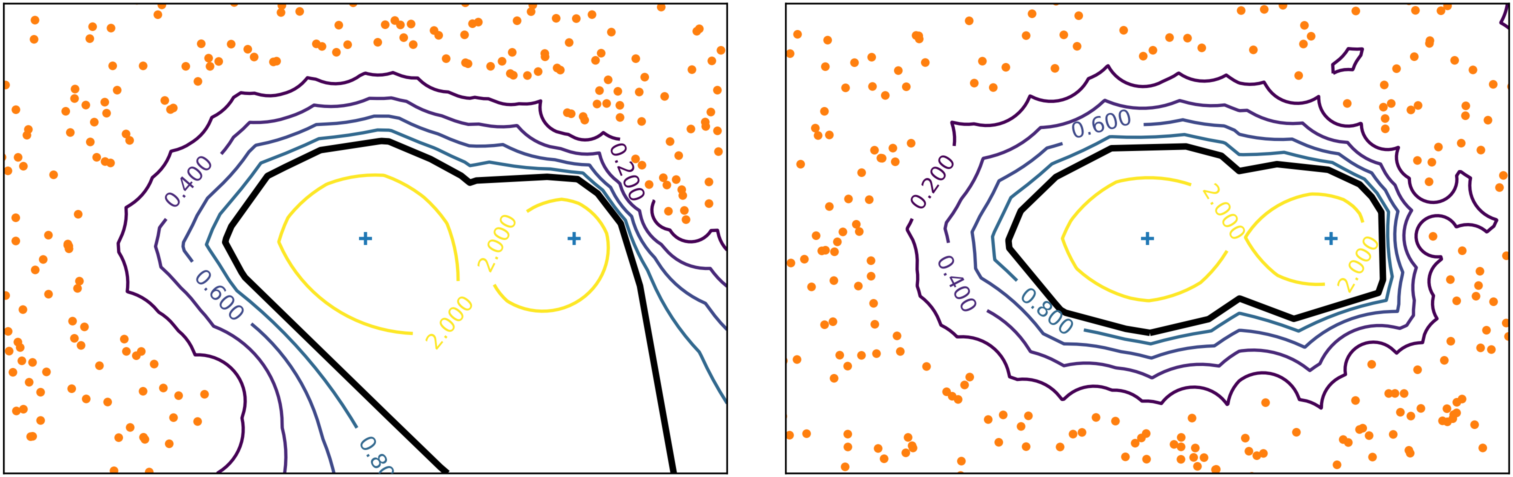
    (intuitively $\gamma \leq 1$ if $+$ is rare)

$$d_\gamma(x, x_i) = \begin{cases} d(x, x_i) & \text{if } x_i \in S_-, \\ \gamma \cdot d(x, x_i) & \text{if } x_i \in S_+. \end{cases}$$

γ-NN: varying γ with two points



γ-NN: varying γ with a few $+$



- γ-NN can control
  how close to the minuses it pushes the boundary

γ-NN: Algorithm

Algorithm 1: Classification of a new example with γk−NN

Input : a query $\mathbf{x}$ to be classified, a set of labeled samples $S = S_+ \cup S_-$, a number of neighbors k, a positive real value γ, a distance function d

Output: the predicted label of $\mathbf{x}$

$\mathcal{NN}^-, \mathcal{D}^- \leftarrow nn(k, \mathbf{x}, S_-)$ // nearest negative neighbors with their distances

$\mathcal{NN}^+, \mathcal{D}^+ \leftarrow nn(k, \mathbf{x}, S_+)$ // nearest positive neighbors with their distances

$\mathcal{D}^+ \leftarrow \gamma \cdot \mathcal{D}^+$

$\mathcal{NN}_\gamma \leftarrow firstK\left(k, sortedMerge((\mathcal{NN}^-, \mathcal{D}^-), (\mathcal{NN}^+, \mathcal{D}^+))\right)$

$y \leftarrow +$ if $\left|\mathcal{NN}_\gamma \cap \mathcal{NN}^+\right| \geq \frac{k}{2}$ else $-$ // majority vote based on $\mathcal{NN}_\gamma$

return y

- Trivial to implement
- Same complexity as k-NN (at most twice)
- Training
  - none, as k-NN
  - γ is selected by cross-validation (on the measure of interest)

γ-NN: a way to reweight distributions

- In uncertain regions
- At the boundaries

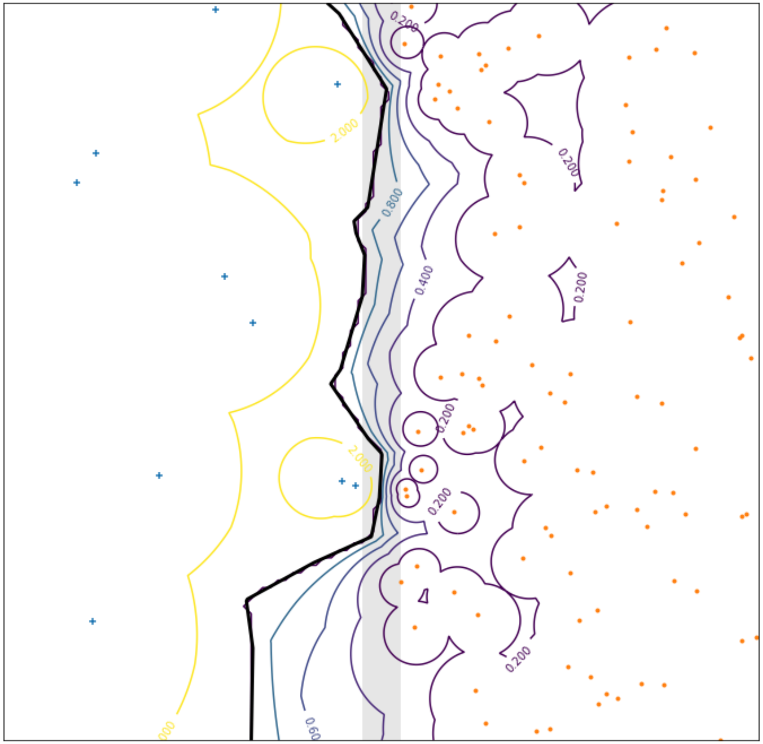
Results on public datasets (F-measure)

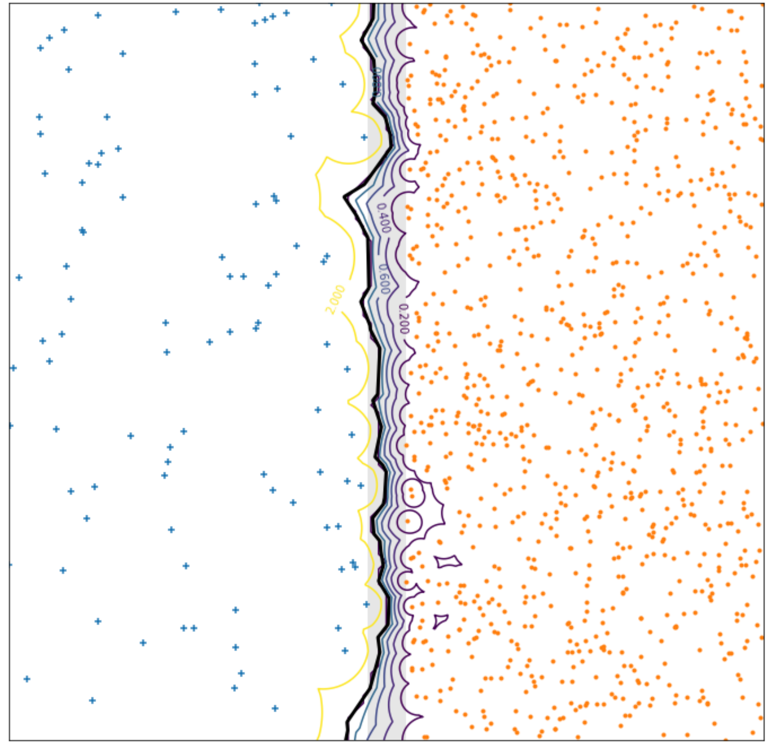
DATASETS	3−NN	DUPk−NN	Wk−NN	CWk−NN	LMNN	γk−NN
BALANCE	0.954(0.017)	0.954(0.017)	0.957(0.017)	0.961(0.010)	**0.963(0.012)**	0.954(0.029)
AUTOMPG	0.808(0.077)	0.826(0.033)	0.810(0.076)	0.815(0.053)	0.827(0.054)	**0.831(0.025)**
IONO	0.752(0.053)	0.859(0.021)	0.756(0.060)	0.799(0.036)	0.890(0.039)	**0.925(0.017)**
PIMA	0.500(0.056)	0.539(0.033)	0.479(0.044)	0.515(0.037)	0.499(0.070)	**0.560(0.024)**
WINE	0.881(0.072)	0.852(0.057)	0.881(0.072)	0.876(0.080)	**0.950(0.036)**	0.856(0.086)
GLASS	0.727(0.049)	0.733(0.061)	0.736(0.052)	0.717(0.055)	0.725(0.048)	**0.746(0.046)**
GERMAN	0.330(0.030)	0.449(0.037)	0.326(0.030)	0.344(0.029)	0.323(0.054)	**0.464(0.029)**
VEHICLE	0.891(0.044)	0.867(0.027)	0.891(0.044)	0.881(0.021)	**0.958(0.020)**	0.880(0.049)
HAYES	0.036(0.081)	0.183(0.130)	0.050(0.112)	0.221(0.133)	0.036(0.081)	**0.593(0.072)**
SEGMENTATION	0.859(0.028)	0.862(0.018)	0.877(0.028)	0.851(0.022)	**0.885(0.034)**	0.848(0.025)
ABALONE8	0.243(0.037)	0.318(0.013)	0.241(0.034)	0.330(0.015)	0.246(0.065)	**0.349(0.018)**
YEAST3	0.634(0.066)	0.670(0.034)	0.634(0.066)	**0.699(0.015)**	0.667(0.055)	0.687(0.033)
PAGEBLOCKS	0.842(0.020)	0.850(0.024)	0.849(0.019)	0.847(0.029)	**0.856(0.032)**	0.844(0.023)
SATIMAGE	0.454(0.039)	0.457(0.027)	0.454(0.039)	0.457(0.023)	**0.487(0.026)**	0.430(0.008)
LIBRAS	**0.806(0.076)**	0.788(0.187)	**0.806(0.076)**	0.789(0.097)	0.770(0.027)	0.768(0.106)
WINE4	0.031(0.069)	**0.090(0.086)**	0.031(0.069)	0.019(0.042)	0.000(0.000)	**0.090(0.036)**
YEAST6	0.503(0.302)	0.449(0.112)	0.502(0.297)	0.338(0.071)	0.505(0.231)	**0.553(0.215)**
ABALONE17	0.057(0.078)	**0.172(0.086)**	0.057(0.078)	0.096(0.059)	0.000(0.000)	0.100(0.038)
ABALONE20	0.000(0.000)	0.000(0.000)	0.000(0.000)	**0.067(0.038)**	0.057(0.128)	0.052(0.047)
MEAN	0.543(0.063)	0.575(0.053)	0.544(0.064)	0.559(0.046)	0.560(0.053)	**0.607(0.049)**

Results on DGFiP datasets (F-measure)

DATASETS	3−NN	γk−NN	SMOTE	SMOTE+γk−NN
DGFIP19 2	0,454 (0,007)	<u>0,528</u> (0,005)	0,505 (0,010)	**0,529** (0,003)
DGFIP9 2	0,173 (0,074)	<u>0,396</u> (0,018)	0,340 (0,033)	**0,419** (0,029)
DGFIP4 2	0,164 (0,155)	<u>0,373</u> (0,018)	0,368 (0,057)	**0,377** (0,018)
DGFIP8 1	0,100 (0,045)	**0,299** (0,010)	0,278 (0,043)	**0,299** (0,011)
DGFIP8 2	0,140 (0,078)	0,292 (0,028)	**0,313** (0,048)	<u>0,312</u> (0,021)
DGFIP9 1	0,088 (0,090)	0,258 (0,036)	<u>0,270</u> (0,079)	**0,288** (0,026)
DGFIP4 1	0,073 (0,101)	<u>0,231</u> (0,139)	0,199 (0,129)	**0,278** (0,067)
DGFIP16 1	0,049 (0,074)	0,166 (0,065)	<u>0,180</u> (0,061)	**0,191** (0,081)
DGFIP16 2	0,210 (0,102)	0,202 (0,056)	<u>0,220</u> (0,043)	**0,229** (0,026)
DGFIP20 3	0,142 (0,015)	<u>0,210</u> (0,019)	0,199 (0,015)	**0,212** (0,019)
DGFIP5 3	0,030 (0,012)	0,105 (0,008)	**0,110** (0,109)	<u>0,107</u> (0,010)
MEAN	0,148 (0,068)	<u>0,278</u> (0,037)	0,271 (0,057)	**0,295** (0,028)

γ-NN at a boundary (10 and 100 +)





Overview

- Introduction
- High-dimensional problems
  - The curse of dimensionality
  - Ockham's Razor
  - Notions of Simplicity
- High-dimensionality and Neighborhood
- Imbalanced classification problems
  - The Problem (and performance measures)
  - Reweight, resampling, etc
  - Correcting k-NN (γ-***NN***)
  - Focusing on the F-Measure optimization (Élisa)
- Discussion

From Cost-Sensitive Classification to Tight F-measure Bounds

- **Kevin Bascol**, Rémi Emonet, Elisa Fromont, Amaury Habrard, **Guillaume Metzler**, Marc Sebban
- AISTATS2019

Optimizing the F_β-measure?

- Reminder
  - Precision: $prec = \dfrac{TP}{TP + FP}$
  - Recall: $rec = \dfrac{TP}{P} = \dfrac{TP}{TP + FN}$
  - F_β-measure: $F_\beta = (1 + \beta^2)\dfrac{prec \cdot rec}{\beta^2 \cdot prec + rec}$
- **Non-separability**, i.e. $F_\beta \neq \sum_{(x_i, y_i) \in S} \dots$

NB: accuracy is separable, $acc = \sum_{(x_i, y_i) \in S} \frac{1}{m} \delta(y_i - \hat{y}_i)$

$\Rightarrow$ The loss for one point depends on the others
$\Rightarrow$ Impossible to optimize directly
$\Rightarrow$ Impossible to optimize on a subset (minibatch)

Weighted classification for F_β

$$F_\beta = \frac{(1+\beta^2)\cdot(P-FN)}{1+\beta^2 P - FN + FP} = \frac{(1+\beta^2)\cdot(P-e_1)}{1+\beta^2 P - e_1 + e_2}$$

- The F_β-measure is linear fractional *(in $e = (e_1, e_2) = (FN, FP)$)*
  i.e. $F_\beta = \dfrac{\langle a', e\rangle + b}{\langle c, e\rangle + d} = \dfrac{A}{B}$
- Relation to weighted classification

$F_\beta \geq t$ (we achieve a good, above t, F_β value)

$$\Leftrightarrow A \geq t \cdot B$$
$$\Leftrightarrow A - t \cdot B \geq 0$$
$$\Leftrightarrow (1+\beta^2)\cdot(P-e_1) - t(1+\beta^2 P - e_1 + e_2) \geq 0$$
$$\Leftrightarrow (-1-\beta^2+t)e_1 - te_2 \geq -P(1+\beta^2) + t(1+\beta^2 P)$$
$$\Leftrightarrow (1+\beta^2-t)e_1 + te_2 \leq -P(1+\beta^2) + t(1+\beta^2 P)$$

⇒ so, we can minimize the weighted problem
with class weights $a(t) = (1+\beta^2-t, t)$

Overview

- Introduction
- High-dimensional problems
  - The curse of dimensionality
  - Ockham's Razor
  - Notions of Simplicity
- High-dimensionality and Neighborhood
- Imbalanced classification problems
  - The Problem (and performance measures)
  - Reweight, resampling, etc
  - Correcting k-NN (γ-***NN***)
  - Focusing on the F-Measure optimization (Élisa)
- <u>Discussion</u>

Thank you! Questions?

and now for something completely different…