The old EM algorithm for quantification learning: Some past and recent results

Marco Saerens (UCLouvain, Belgium),
Christine Decaestecker (ULB, Belgium)

ECML PKDD 2022

UCLouvain ULB

Table of contents

- Introduction: initial case study
- Intuitive derivation of the EM algorithm for prior probability shift
- Interesting lessons from some recent advances

2

Introduction: Initial case study

3

Motivation

- The work was published in the early 2000s
  - Saerens M. Decaestecker C. & Latinne P. (2001). “Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure”. Neural computation, 14 (1), pp. 21-41.
- We were confronted to the following challenging problem
  - To classify pixels of images
  - Based on remote sensing information
  - = To provide a Land cover interpretation

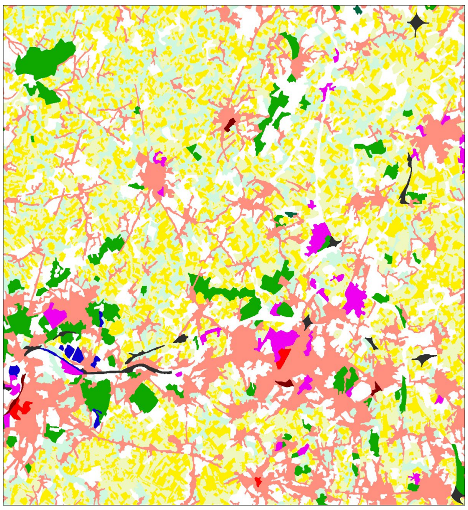
4

Motivations

- Real data coming from
  - LANDSAT Thematic Mapper 7 bands
  - 36km x 36km
  - 1201 x 1201 « pixels » to classify
  - 11 class labels
  - 50 features from spectral/textural filters

5

Motivation



6

Motivations

- Some of the 11 classes
  - Arable, cultivated, land
  - Road network
  - Industrial, commercial unit
  - Forest
  - Urban fabric
  - …

7

Motivations

- The problem is
  - Strongly unbalanced
  - Class priors (prevalence) vary from one map to another!
- It means that a classification model
  - Trained on one map
  - Is not suited when applied on another map
  - Because class priors differ (prior shift)

8

Motivations

- Three main ideas emerged from this challenging problem
  - Use unlabeled data from the test set in order to improve the classification model
  - Try to adapt an already existing classification model to new conditions
  - Estimate the a priori probabilities in new conditions

9

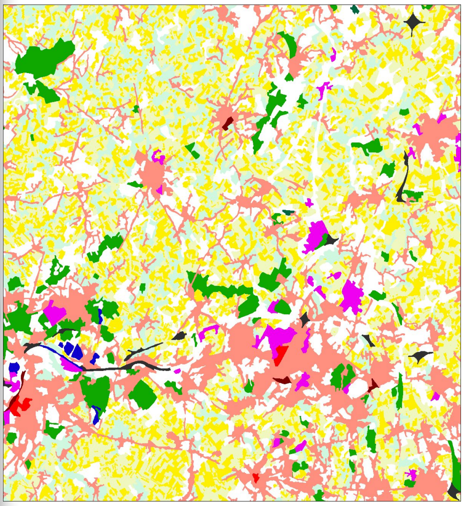
Motivations

- Both idea were largely exploited during this period (end nineties and beginning of the 2000s)
  - Semi-supervised classification
  - Transfer learning (prior shift, label shift, etc)

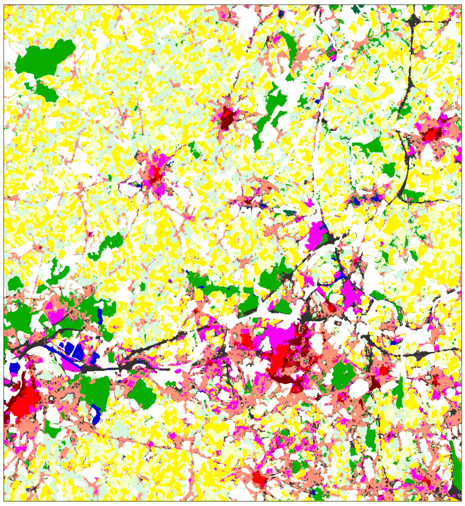
10

Motivations

Reference, test, image

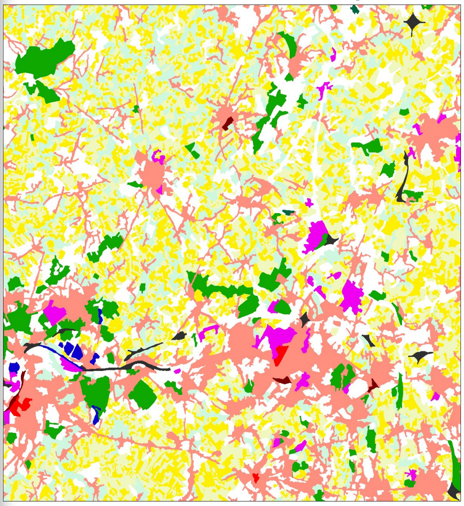


Bagging classification model

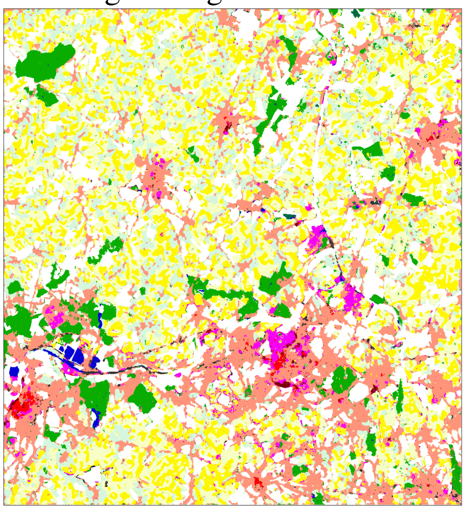


Motivations

Reference, test, image



Logistic regression + EM

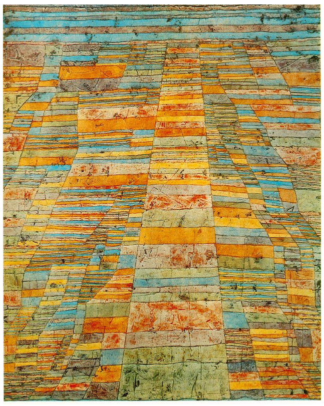


Definition of the problem

- How can we adapt a classification model to new a priori probability conditions?
  - When the new a priori probabilities are known
  - When these new a priori probabilities are unknown

13

The Expectation-Maximization algorithm



14

Definition of the problem

- This EM technique (Latinne et al., 2001; Saerens et al., 2001) is also called the
  - Maximum likelihood method or
  - The iterative label or prior shift correction

15

Definition of the problem

- Assume we have some calibrated classification model providing
  - exact a posteriori probabilities of membership to a set of q classes $\{\omega_i\}_{i=1}^{q}$
  - based on some observed feature vector $\mathbf{x}$ for the random vector $\boldsymbol{x}$, simply denoted as

$$\mathrm{P}(\omega_i|\mathbf{x}) = \mathrm{P}(y = \omega_i|\boldsymbol{x} = \mathbf{x})$$

  - This is a kind of "perfect model matching" assumption

16

Dealing with changing a priori probabilities: priors known

- This classifier provides a posteriori probabilities

$$\mathrm{P}_t(y = \omega_i | \mathbf{x})$$

  - In the conditions of the training set (subscript t)
  - Assume that we know the priors of both training and test ("real life") sets

$$\mathrm{P}_t(\omega_i) = \mathrm{P}_t(y = \omega_i), \text{ and } \mathrm{P}(\omega_i) = \mathrm{P}(y = \omega_i)$$

  - which do not match (training prior ≠ "real life" prior):

$$\begin{cases} \mathrm{P}_t(y = \omega_i) \neq \mathrm{P}(y = \omega_i) \\ \mathrm{P}_t(\mathbf{x}|y = \omega_i) = \mathrm{P}(\mathbf{x}|y = \omega_i) \end{cases}$$

17

Dealing with changing a priori probabilities: priors known

- We are seeking the a posteriori probabilities in the conditions of the real-life dataset (no subscript t)

$$\mathrm{P}(y = \omega_i | \mathbf{x})$$

- We have from Bayes' rule

$$\begin{cases} \mathrm{P}_t(\mathbf{x}|y = \omega_i) = \dfrac{\mathrm{P}_t(y = \omega_i|\mathbf{x})\, \mathrm{P}_t(\mathbf{x})}{\mathrm{P}_t(y = \omega_i)} \\ \mathrm{P}(\mathbf{x}|y = \omega_i) = \dfrac{\mathrm{P}(y = \omega_i|\mathbf{x})\, \mathrm{P}(\mathbf{x})}{\mathrm{P}(y = \omega_i)} \end{cases}$$

18

Dealing with changing a priori probabilities: priors known

- Thus

$$\frac{\mathrm{P}(y=\omega_i|\mathbf{x})\,\mathrm{P}(\mathbf{x})}{\mathrm{P}(y=\omega_i)} = \frac{\mathrm{P}_t(y=\omega_i|\mathbf{x})\,\mathrm{P}_t(\mathbf{x})}{\mathrm{P}_t(y=\omega_i)}$$

- From which we isolate the posteriors in real-life (test) conditions, $\mathrm{P}(y=\omega_i|\mathbf{x})$

19

Dealing with changing a priori probabilities: priors known

- We easily obtain

$$f(\mathbf{x}) = \frac{\mathrm{P}_t(\mathbf{x})}{\mathrm{P}(\mathbf{x})}$$

$$\begin{aligned}\mathrm{P}(y=\omega_i|\mathbf{x}) &= \frac{\mathrm{P}_t(\mathbf{x})}{\mathrm{P}(\mathbf{x})}\frac{\mathrm{P}_t(y=\omega_i|\mathbf{x})\,\mathrm{P}(y=\omega_i)}{\mathrm{P}_t(y=\omega_i)} \\ &= f(\mathbf{x})\frac{\mathrm{P}_t(y=\omega_i|\mathbf{x})\,\mathrm{P}(y=\omega_i)}{\mathrm{P}_t(y=\omega_i)} \\ &= f(\mathbf{x})\,\mathrm{P}_t(y=\omega_i|\mathbf{x})\,\mathrm{odds}(y=\omega_i)\end{aligned}$$

$$\mathrm{odds}(y=\omega_i) = \frac{\mathrm{P}(y=\omega_i)}{\mathrm{P}_t(y=\omega_i)}$$

(= weighting factor common in sampling theory)

20

Dealing with changing a priori probabilities: priors known

- But since

$$\sum_{i=1}^{q} \mathrm{P}(y = \omega_i | \mathbf{x}) = 1$$

- we have

$$f(\mathbf{x}) = \left[\sum_{i=1}^{q} \mathrm{P}_t(y = \omega_i | \mathbf{x}) \, \mathrm{odds}(y = \omega_i) \right]^{-1}$$

21

Dealing with changing a priori probabilities: priors known

- We thus obtain the "new" a posteriori probabilities for the real-life, test, data

$$\mathrm{P}(y = \omega_i | \mathbf{x}) = \frac{\mathrm{P}_t(y = \omega_i | \mathbf{x}) \dfrac{\mathrm{P}(\omega_i)}{\mathrm{P}_t(\omega_i)}}{\displaystyle\sum_{j=1}^{q} \mathrm{P}_t(y = \omega_j | \mathbf{x}) \dfrac{\mathrm{P}(\omega_j)}{\mathrm{P}_t(\omega_j)}}$$

22

Dealing with changing a priori probabilities: priors unknown

- Now, intuitively, if the priors are not known in advance (sattelite image classification), iterate on all samples of the test set:
  1. Estimate the new a priori probabilities based on the adjusted results of the classifier on the real-world data set
     $$\mathrm{P}(\omega_i) = \frac{1}{n}\sum_{k=1}^{n} \mathrm{P}(y_k = \omega_i|\mathbf{x}_k)$$
  2. Re-estimate the a posteriori probabilities based on the current estimates of the a priori probabilities
     $$\mathrm{P}(y_k = \omega_i|\mathbf{x}_k) = \frac{\mathrm{P}_t(y_k = \omega_i|\mathbf{x}_k)\dfrac{\mathrm{P}(\omega_i)}{\mathrm{P}_t(\omega_i)}}{\sum_{j=1}^{q}\mathrm{P}_t(y_k = \omega_j|\mathbf{x}_k)\dfrac{\mathrm{P}(\omega_j)}{\mathrm{P}_t(\omega_j)}}$$

23

Dealing with changing a priori probabilities: priors unknown

- This was reformulated as an instance of the EM algorithm
  – maximizing the log-likelihood of the real data sample

- The method is an easy-to-implement post-processing technique

24

Some recent advances

More recent results

- We investigated recent papers published
  - In major conference proceedings
  - In major journals
  - The list is certainly not comprehensive though
- It appears that both
  - The "Adjusted classify and count" method (Forman, 2005, 2006)
  - The "EM" algorithm
- are still studied and in use, often as baselines

More recent results

- This is probably due to two factors
  - The arise of the fields of "transfer learning"
  - as well as "learning to quantify"
- Note that the idea behind quantification
  - Already appeared in the biomedical field (epidemiology, etc) long ago
  - As well as in pattern recogniton (see, e.g., McLachlan, 1992)

27

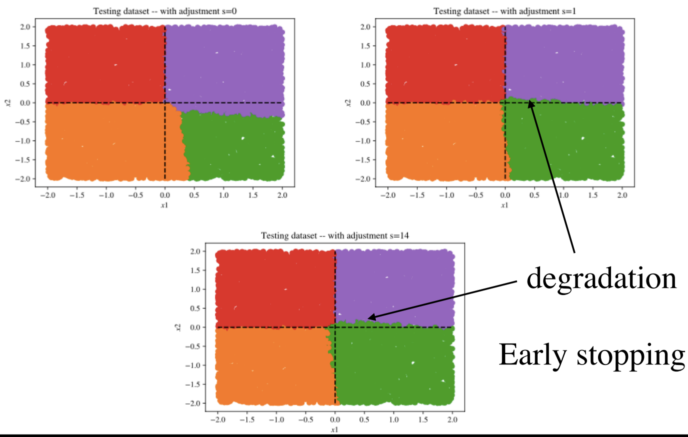
Lessons from recent results: first lesson

- In practice, there exists probably more stable algorithms than the EM
  - Indeed, du Plessis et al. (2014) and Alexandari et al. (2020) showed that
  - the corresponding optimization problem is concave
- However, the EM can get stuck in a degenerate fix point (du Plessis et al., 2014) !
  - Indeed, a posteriori probability vector putting all observations in the same class is a fix point of the EM !

28

Lessons from recent results: first lesson

- In addition, the EM sometimes exaggerates the adjustments (Caelen, 2018)



Early stopping rule??

Lessons from recent results: first lesson

- The priors can be computed by maximizing a concave function (the likelihood of the test set)
  - Indeed following (du Plessis et al., 2014; Alexandari, 2020); see also (Tasche, 2017),
  - Assuming an iid sample,
- The likelihood of the test set is:

$$\prod_{k=1}^{n} \mathrm{P}(\boldsymbol{x}_k = \mathbf{x}_k)$$

  - Let's calculate this quantity

$$
\begin{aligned}
\prod_{k=1}^{n} \mathrm{P}(\boldsymbol{x}_k = \mathbf{x}_k) &= \prod_{k=1}^{n} \sum_{i=1}^{q} \mathrm{P}(y_k = \omega_i, \boldsymbol{x}_k = \mathbf{x}_k) \\
&= \prod_{k=1}^{n} \sum_{i=1}^{q} \mathrm{P}(\mathbf{x}_k | y_k = \omega_i) \mathrm{P}(y_k = \omega_i) \\
&= \prod_{k=1}^{n} \sum_{i=1}^{q} \mathrm{P}_t(\mathbf{x}_k | y_k = \omega_i) \mathrm{P}(y_k = \omega_i) \\
&= \prod_{k=1}^{n} \sum_{i=1}^{q} \frac{\mathrm{P}_t(y_k = \omega_i | \mathbf{x}_k) \mathrm{P}_t(\mathbf{x}_k)}{\mathrm{P}_t(y_k = \omega_i)} \mathrm{P}(y_k = \omega_i) \\
&= \prod_{k=1}^{n} \mathrm{P}_t(\mathbf{x}_k) \sum_{i=1}^{q} \frac{\mathrm{P}_t(y_k = \omega_i | \mathbf{x}_k)}{\mathrm{P}_t(y_k = \omega_i)} \mathrm{P}(y_k = \omega_i) \\
&= \prod_{k=1}^{n} \mathrm{P}_t(\mathbf{x}_k) \sum_{i=1}^{q} \frac{\mathrm{P}_t(y_k = \omega_i | \mathbf{x}_k)}{\mathrm{P}_t(\omega_i)} \mathrm{P}(\omega_i) \\
&= \left(\prod_{k=1}^{n} \mathrm{P}_t(\mathbf{x}_k) \right) \times \left(\prod_{k=1}^{n} \sum_{i=1}^{q} \frac{\mathrm{P}_t(y_k = \omega_i | \mathbf{x}_k)}{\mathrm{P}_t(\omega_i)} \mathrm{P}(\omega_i) \right)
\end{aligned}
$$

31

Lessons from recent results: first lesson

- Taking the log of the likelihood provides

$$
\sum_{k=1}^{n} \log \mathrm{P}_t(\mathbf{x}_k) + \sum_{k=1}^{n} \log \sum_{i=1}^{q} \frac{\mathrm{P}_t(y_k = \omega_i | \mathbf{x}_k)}{\mathrm{P}_t(\omega_i)} \mathrm{P}(\omega_i)
$$

- Finally, we have to maximize the following concave objective function with respect to the priors

$$
\sum_{k=1}^{n} \log \left(\sum_{i=1}^{q} \frac{\mathrm{P}_t(y_k = \omega_i | \mathbf{x}_k)}{\mathrm{P}_t(\omega_i)} \mathrm{P}(\omega_i) \right)
$$

subject to $\mathrm{P}(\omega_i) \geq 0$ and $\sum_{i=1}^{q} \mathrm{P}(\omega_i) = 1$

32

Lessons from recent results: first lesson

- So, why not directly maximize this concave function?
  - This is what was recently exploited by Alexandari et al. (2020), as well as Sipka et al. (2022) based on the confusion matrix
  - The objective function is very close to the log-likelihood of finite mixture models, also containing the priors(1)

(1) Note: just after the presentation, we noticed the following. From (McLachlan, 2000, section 2.8), the application of the EM to maximize this objective function seems to provide the same equations as the EM algorithm of Saerens et al. (2001). This is still to be verified, though.

33

Lessons from recent results: first lesson

- This rises some remarks/questions like
  - Does the maximization of the concave objective function provide the same solution as the EM?
  - Can we find an efficient procedure for computing the maximum of this objective function?

34

Lessons from recent results: first lesson

- Moreover, du Plessis et al. (2014) further showed that
  - The EM algorithm is equivalent to Kullback-Leibler divergence between train likelihood and test likelihood
  - It also proposes a technique for approximating new priors in the more general case of f-divergences

35

Lessons from recent results: first lesson

- It was also shown by Tasche (2017) that both
  - the “Adjusted classify and count” technique and
  - the “EM” technique
- are Fisher consistent
  - This is a desirable property of an estimator, in the same spirit as unbiasedness or asymptotic consistency

36

Lessons from recent results: first lesson

- Note that the same author (Tasche, 2022) recently extended the EM method to
  - prior probability + covariate shifts
  - by making some factorization assumptions
- The EM algorithm (and also the adjusted classify and count) has recently been extended in order to deal with ordinal data (Bunse, 2022)

37

Lessons from recent results: second lesson

- Calibration of the classification model is essential
  - Let us consider a binary classification problem with target variable $y = 0, 1$
  - Denote by $g(\mathbf{x})$ the probabilistic output (soft prediction) of the classification model for feature vector $\mathbf{x}$
- Then, intuitively, the classification model is perfectly calibrated on a domain D of the feature space when

$$\hat{y} \triangleq g(\mathbf{x}) = \mathbb{E}[y|\boldsymbol{x} = \mathbf{x}] \text{ for all } \mathbf{x} \in \mathcal{D}$$

  - That is, the output of the model matches true posteriors

38

Lessons from recent results: second lesson

- But since this is difficult to verify in practice for all $\mathbf{x}$, we often simply require (e.g., De Groot, 1983)

$$\hat{y} = \mathbb{E}[y|g(\mathbf{x})] \text{ for all } g(\mathbf{x}) \in [0, 1]$$

- The importance of calibration has been highlighted in several recent works,
  - Recently by (Alexandari et al., 2020; Garg et al., 2020; Esuli et al., 2021)
- Calibration looks important
  - Not only for quantification, but also for interpretability (Scafarto, 2022)

39

Lessons from recent results: second lesson

- But when is a classification model well-calibrated?
- It depends on multiple factors! Among which:
  - The model has the "perfect model matching" property
  - The training set is unbiased
  - The minimum of the cost function is reached (model well-fitted)
  - The cost function for training the model minimizes to the conditional expectation

40

Lessons from recent results: second lesson

- Calibration is often performed by using a post-processing step (Guo, 2017; Alexandari et al., 2020; Garg et al., 2020)
  - Involving a validation set
  - Many deep learning models have a competitive classification accuracy but are often ill-calibrated (Guo, 2017)!
- But other avenues could be explored
  - For instance, considering the cost function

41

Lessons from recent results: second lesson

- ML researchers (e.g., Hampshire, 1990) studied the conditions under which the minimum of the cost function is the conditional expectation
  - This is closely related to the study of proper scoring rules in applied statistics (see, e.g., De Groot, 1983; Gneiting, 2007)
- Under some mild assumptions, for binary classification, the condition (Hampshire, 1990) is

$$\frac{(\widehat{y}-1)}{\widehat{y}} = \frac{\pounds'[\widehat{y};1]}{\pounds'[\widehat{y};0]}$$

  - where $\pounds$ is the loss associated to each observation

42

Lessons from recent results: second lesson

- This condition is also sufficient
- These results generalize to q classes

Lessons from recent results: second lesson

- For the least square error criterion

$$\pounds[\widehat{y}; y] = \frac{1}{2}(\widehat{y} - y)^2$$
$$\pounds'[\widehat{y}; y] = (\widehat{y} - y)$$

- The derivatives are

$$\begin{cases} \pounds'[\widehat{y}; 1] = (\widehat{y} - 1) \\ \pounds'[\widehat{y}; 0] = \widehat{y} \end{cases}$$

– and the condition is fulfilled

Lessons from recent results: second lesson

- For the log-likelihood ("cross-entropy") criterion

$$\pounds[\widehat{y}; y] = y \ln(\widehat{y}) + (1-y) \ln(1-\widehat{y})$$

- Exercice: Does the log-likelihood criterion lead to the estimation of a posteriori probabilities?

- Questions:
  – In deep learning, which cost functions (Katarzyna et al., 2016) minimize to a posteriori probabilities?
  – What are the empirical consequences of this?

Lessons from recent results: second lesson

- In addition, it has also been shown under some assumptions that (Lindley, 1982; Saerens et al., 2002)
  – If the classification model has been trained with an arbitrary cost function and this cost function is minimized
  – There exists a transformation mapping the model's predictions to a posteriori probabilities

- This transformation is $f(\hat{y}) = \dfrac{1}{1 - \dfrac{\pounds'(\hat{y};1)}{\pounds'(\hat{y};0)}}$
  – This can also be generalized to q classes

Lessons from recent results: second lesson

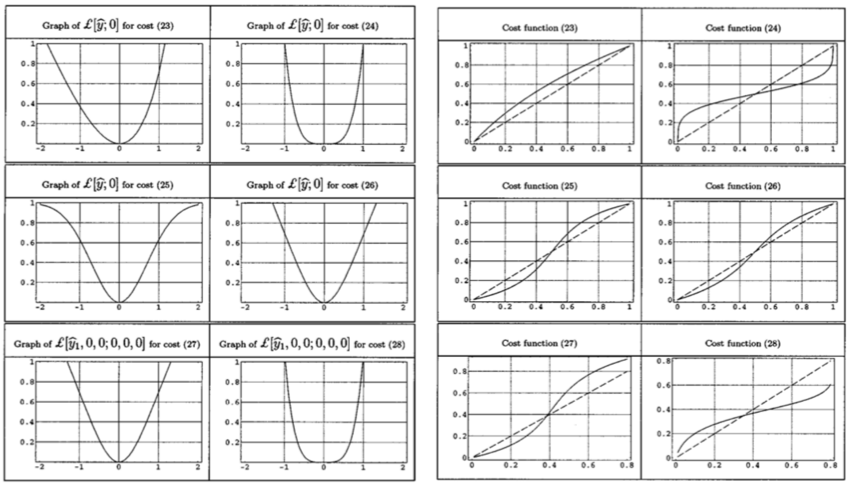
- Here is an example with six different loss functions

$$\pounds[\hat{y};y] = \exp[y]\,(y - \hat{y} - 1) + \exp[\hat{y}] \quad (23)$$
$$\pounds[\hat{y};y] = (\hat{y} - y)^4 \quad (24)$$
$$\pounds[\hat{y};y] = 1 - \exp[-(\hat{y} - y)^2] \quad (25)$$
$$\pounds[\hat{y};y] = \log[1 + (\hat{y} - y)^2] \quad (26)$$
$$\pounds[\hat{\mathbf{y}};\mathbf{y}] = \log[1 + \|\hat{\mathbf{y}} - \mathbf{y}\|^2] \quad (27)$$
$$\pounds[\hat{\mathbf{y}};\mathbf{y}] = \exp[\|\hat{\mathbf{y}} - \mathbf{y}\|^2] + \exp[-\|\hat{\mathbf{y}} - \mathbf{y}\|^2] - 2. \quad (28)$$

47

Lessons from recent results: second lesson

- Here are the corresponding remappings (taken from Saerens et al., 2002)



48

Lessons from recent results: second lesson

- Finally, it is of course always useful to represent graphically the predicted values in terms of the observed values
- And use reliability diagrams (see, e.g., Vaicenavicius, 2019)

49

References

- Alexandari A. et al. (2020). “Maximum likelihood with bias-corrected calibration is hard to beat in label shift adaptation”. Proceedings of the 37th International Conference on Machine Learning (ICML 2020), pp. 222-232.
- Bunse M. et al. (2022). “Ordinal quantification through regularization”. Proceedings of the European Conference on Machine Learning (ECML 2022).
- Azizzadenesheli K. et al. (2019). “Regularized learning for domain adaptation under label shift”. Proceedings of the International Conference on Learning Representation (ICLR 2019).
- Bartlett P. et al. (2006). “Convexity, classification, and risk bounds”. Journal of the American Statistical Association, 101 (473), pp. 138-156.
- Caelen O. (2018). “Quantification and learning algorithms to manage prior probability shift”. Master thesis, Institute of Statistics, UCLouvain. Supervisor: M. Saerens; reader: J. Seghers.
- De Groot M. et al. (1983). “The comparison and evaluation of forecasters”. The Statistician, 32, pp. 12-22.
- du Plessis C. et al. (2014). “Semi-supervised learning of class balance under class-prior change by distribution matching”. Neural Networks, 50, pp. 110-119.
- Esuli A. et al. (2021). “A critical reassessment of the Saerens-Latinne-Decaestecker algorithm for posterior probability adjustment”. ACM Transactions on Information Systems, 39 (2), 2.
- Forman G. (2005). “Counting positives accurately despite inaccurate classification”. Proceedings of the European Conference on Machine Learning (ECML 2005), pp. 564-575

50

References

- Forman G. (2006). “Quantifying trends accurately despite classifier error and class imbalance”. Proceedings of the 12th ACM International Conference on Knowledge discovery and Data Mining (KDD 2006), pp. 157-166.
- Garg S. et al. (2020). “A unified view of label shift estimation”. Proceedings of the 34th International Conference in Neural Information Processing Systems (NeurIPS 2020), pp. 3290-3300.
- Gneiting T. et al. (2007). “Strictly proper scoring rules, prediction, and estimation”. *Journal of the American statistical Association*, *102* (477), pp. 359-378.
- Gonzalez P. et al. (2017). “A review on quantification learning”. ACM Computing Surveys, 50 (5), pp. 74: 1-40.
- Guo C. et al. (2017). “On calibration of modern neural networks”. Proceedings of the 34th International Conference on Machine Learning, pp. 1321-1330.
- Hampshire J. et al. (1990). “Equivalence proofs for multi-layer perceptron classifiers and the Bayesian discriminant function”. Proceedings of the 1990 Connectionist Models Summer School, D. Touretzky, J. Elman, T. Sejnowski, and G. Hinton, Eds. San Mateo, CA, pp. 159-172.
- Katarzina J. et al. (2016). “On loss functions for deep neural networks in classification”. Schedae Informaticae (TFML 2017), 25, pp. 49-59.
- Latinne P. et al. (2001). “Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: evidence from a multi-class problem in remote sensing”. Proceedings of the 18th International Conference on Machine Learning (ICML 2001), pp. 298-305.

References

- Lindley D. (1982). “Scoring rules and the inevitability of probabilities”. International Statistical Review, 50, pp. 1-26.
- Lipton Z. et al. (2018). “Detecting and correcting for label shift with black box predictors”. Proceedings of the 35th International Conference on Machine Learning (ICML 2018), pp. 3122- 3130.
- McLachlan G. (1992). “Discriminant analysis and statistical pattern recognition”. Wiley.
- McLachlan G. et al. (2000). “Finite mixture models”. Wiley.
- Moreo, A. et al. (2022). “Tweet sentiment quantification: An experimental re-evaluation”. ArXiv preprint arXiv:2011.08091, pp. 1-21.
- Saerens M. et al. (2001). “Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure”. Neural computation, 14 (1), pp. 21-41.
- Saerens M. et al. (2002). “Any reasonable cost function can be used for a posteriori probability approximation”. IEEE Transactions on Neural Networks, 13 (5), pp. 1204-1210.
- Scafarto G. et al. (2022). “Calibrate to interpret”. Proceedings of the European Conference on Machine Learning (ECML 2022).
- Sipka T. et al. (2022). The hitchhicker’s guide to prior shift adaptation”. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1516-1524.

References

- Tasche D. (2014). “Exact fit of simple finite mixture models”. Journal of Risk and Financial Management, 7 (4), pp. 150-164.
- Tasche D. (2017). “Fisher consistency for prior probability shift”. The Journal of Machine Learning Research, 18 (1), pp. 3338-3369.
- Tasche D. (2022). “Factorisable joint shift in multinomial classification”. Machine Learning and Knowledge Extraction, 4, pp. 779-802.
- Tian J. et al. (2020). “Posterior re-calibration for imbalanced datasets”. Proceedings of the 34th International Conference in Neural Information Processing Systems (NeurIPS 2020), pp. 8101-8113.
- Vaicenavicius J. et al. (2019). “Evaluating model calibration in classification”. Proceedings of the 22th International Conference on Artificial Intelligence and Statistics (AISTATS 2019), pp. 3459-3467.

Thank you for your attention !!