Machine Learning in the wild; from the SPHERE project perspective

Miquel Perelló Nieto 26/04/2022

https://www.bestofbristol.co/12-incredible-aerial-shots-of-bristol/

Training in the lab, deployment in the real world

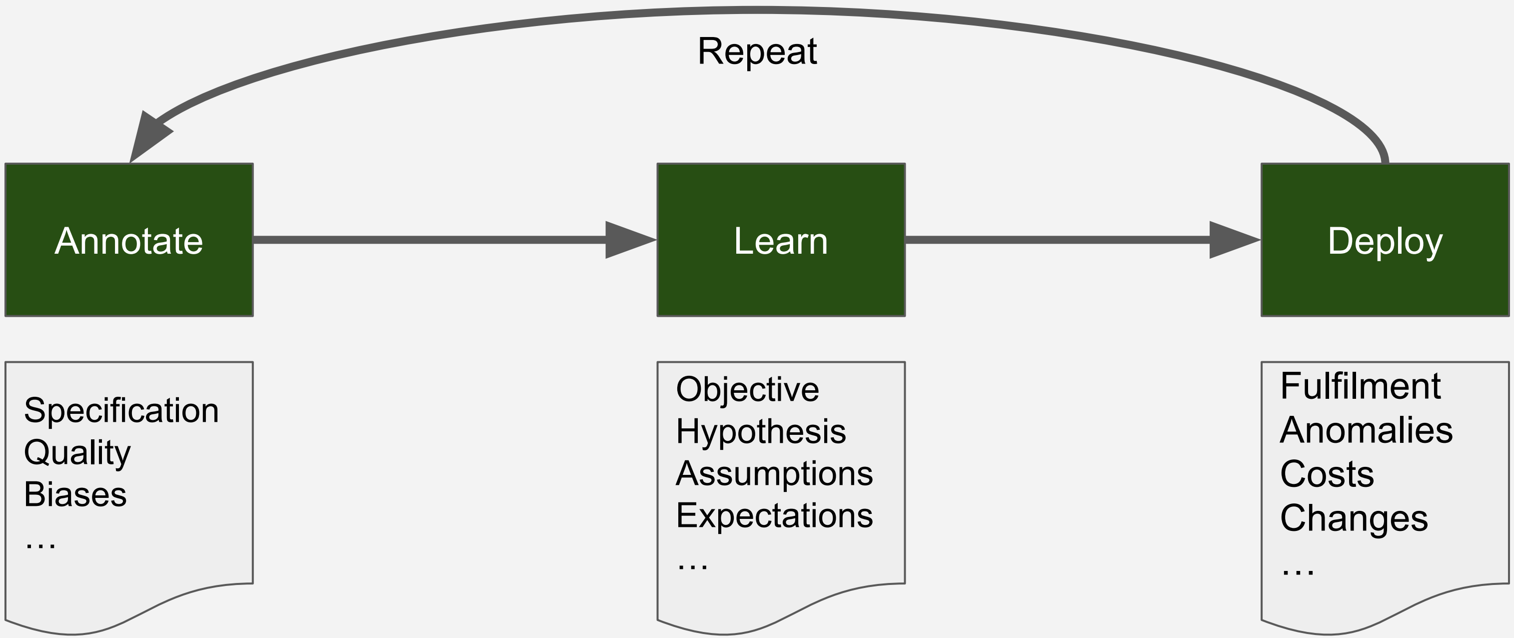
https://www.youtube.com/watch?v=qj9__tNCSOs


https://www.bestofbristol.co/12-incredible-aerial-shots-of-bristol/

SPHERE project summary

- Aims:
  - Monitor free living to detect medical conditions that can not be measured at the hospital
- Objectives:
  - **Collecting** big data from family homes
  - **Compare** the control group against specific conditions
  - **Detect** the specific conditions from the data
- The lab: SPHERE home
  - House with the SPHERE sensors since 2013
  - Controlled environment
  - Wristbands with acceleration and RSSI
  - Sensors in the walls for motion, light, temperature, humidity, human silhouette, water usage, electricity…
- The real world: Bristol
  - More than 50 family homes as a control group
  - Homes with specific conditions: heart valve, hip and knee replacement, Parkinson’s disease, Alzheimer.

Table of content: ML pipeline

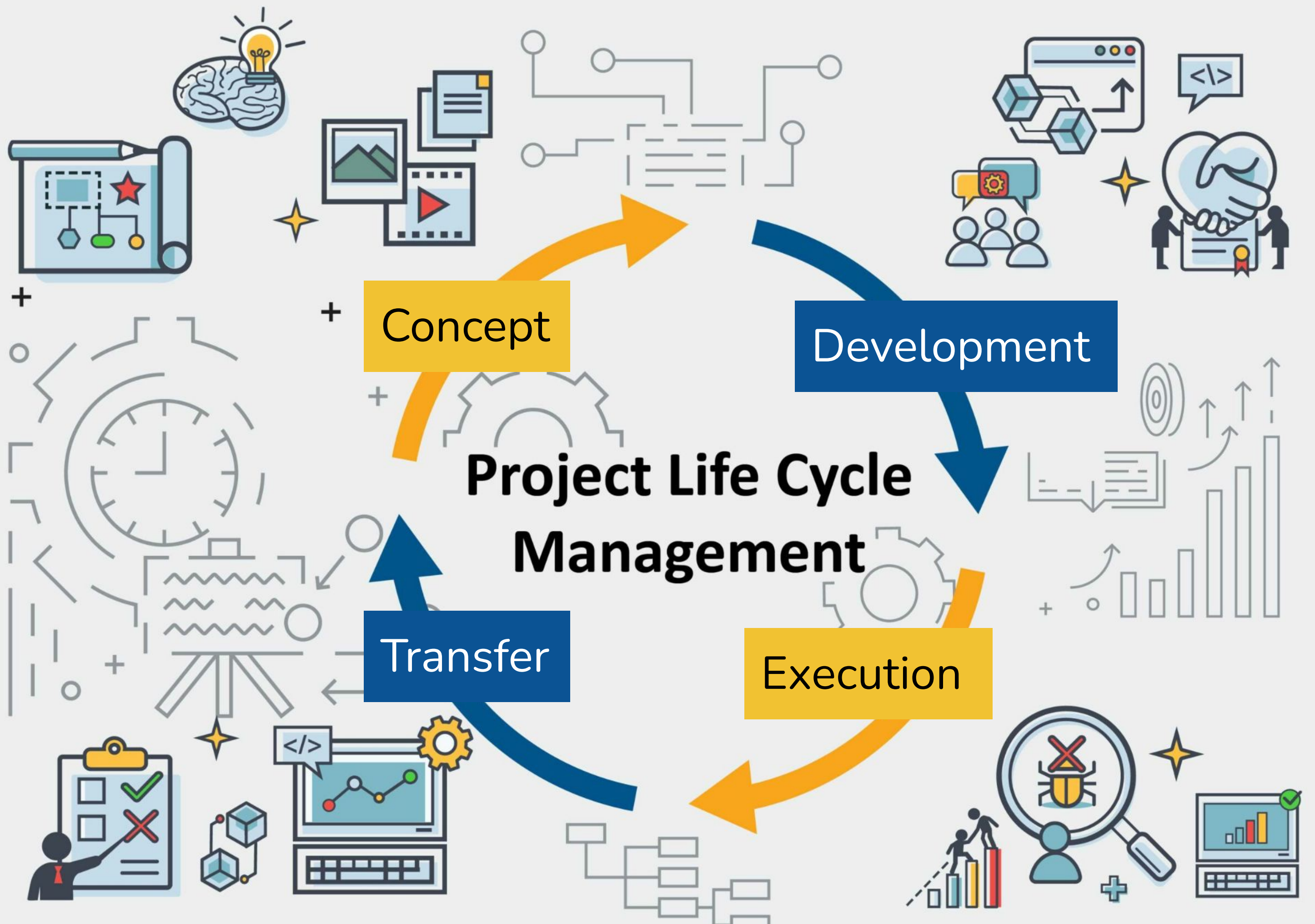


Annotations

Human activity and indoor localisation

- Activities may generalize to certain demographics (but not all!)
- Indoor location fingerprints are different per house

https://www.vecteezy.com/vector-art/2186373-set-of-daily-routines-the-concept-of-daily-life-everyday-leisure-and-work-activities-flat-vector-illustration

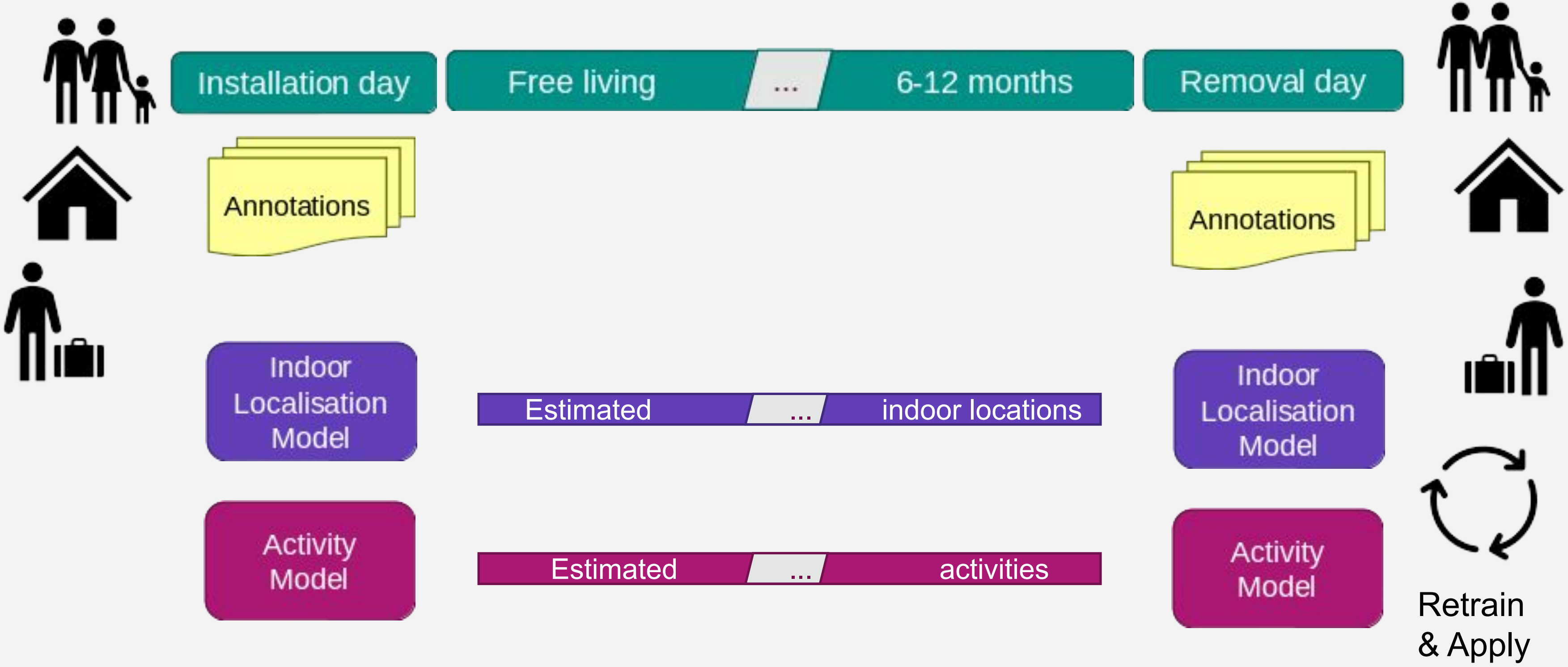


https://www.productdossier.com/blog/why-project-life-cycle-management-is-important/

Concurrency of annotations and project cycle

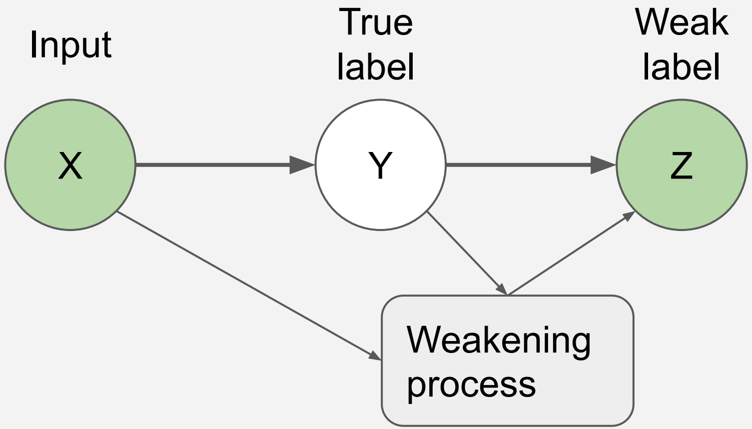
- Sensors and people change over time, while the annotations are concurrently obtained (changes in feature representation)
- Recruited participants may not reflect the population of the conditions of interest (stand still with Parkinson's disease)
- Multiple modalities of annotation with different quality
    - Annotation mistakes
    - Trained technicians
    - Participants with a phone app
    - Participants with pen and paper
    - Post-hoc real time video observation
    - Pseudo labels (e.g. ML generated, or deduction)

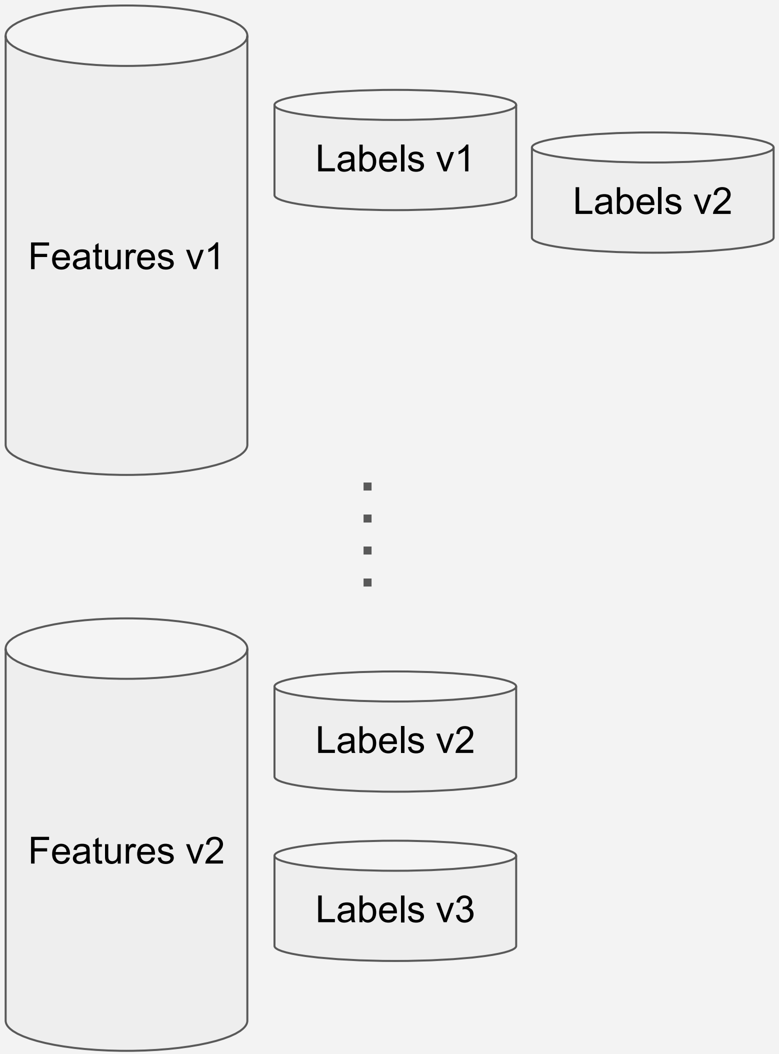
Annotated vs unannotated (1 hour vs 365 days)



Resulting dataset

- Annotations of different quality (including modalities, label sets and noise)
- Sparse annotations (unsupervised)
- Missing or drifting features
- Annotations biased by demographics





Learning

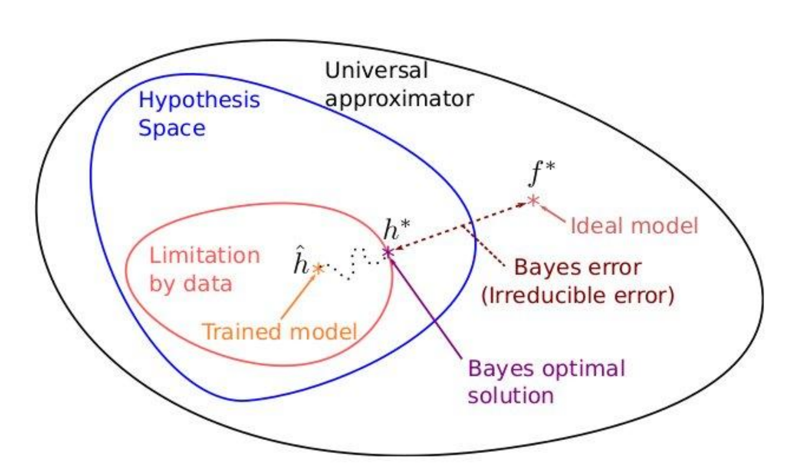
Use the true (and weak) labels to train a model

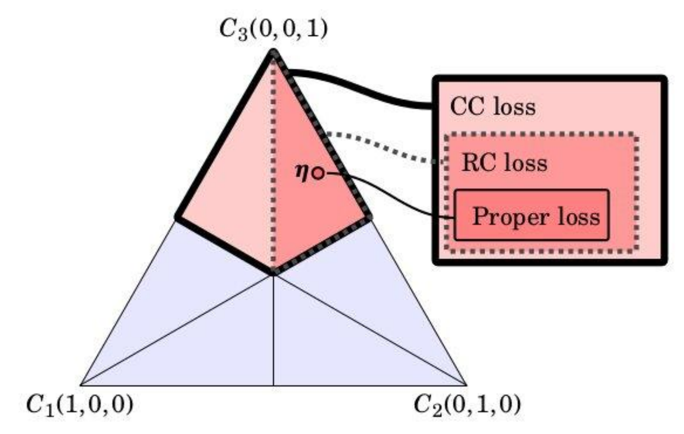
- We have a limited set of annotated activities and locations
- Some activities may generalise (e.g. sitting on the floor)
- Some labels may be weak

https://www.vecteezy.com/vector-art/2186373-set-of-daily-routines-the-concept-of-daily-life-everyday-leisure-and-work-activities-flat-vector-illustration

Types of classifiers and losses

- Class estimators (class calibrated loss)
- Ranking estimators (ranking calibrated loss)
- Score surrogate estimators
- Probability estimators (proper loss)
- Probability estimators with uncertainty
- Label sets (not covered here)





Empirical risk minimization

- Choose the hypothesis that minimizes the expected risk in our training data

$$R_{emp}(h) = \frac{1}{N} \sum_{i=1}^{N} \text{loss}(h(x_i), y_i)$$

$$\hat{h} = \arg\min_{h \in \mathcal{H}} R_{emp}(h)$$

- Assumes i.i.d. data during training and deployment
- Each sample has the same importance (same costs)
- Each class has importance relative to its occurrence (prior distribution)
- In Classification the Bayes optimal minimizes the 0-1 loss
- It is possible to reweight the loss for different costs or priors

Learning with weak labels

- We assume the weakening process can be modeled

$$P(\tilde{\mathrm{Y}} = \tilde{\mathrm{C}}_j | X = \mathbf{x}) = \sum_{i=1}^{K} P(\tilde{\mathrm{Y}} = \tilde{\mathrm{C}}_j | Y = C_i, X = \mathbf{x}) P(Y = C_i | X = \mathbf{x}).$$

- This sum of product can be computed as a matrix multiplication

$$\tilde{\mathbf{q}}(\mathbf{x}) = \mathbf{M}(\mathbf{x})\mathbf{q}(\mathbf{x}),$$

- A common assumption is that matrix **M** does not depend on X (**M(x)->M**)
- Given a known mixing process **M** we can obtain the posterior probabilities for the true class with its pseudoinverse

$$\mathbf{q}(\mathbf{x}) = \mathbf{M}^{+} \tilde{\mathbf{q}}(\mathbf{x}).$$

Losses for weak labels: with known weakening process

Theorem 2.3 ((Jesus Cid-Sueiro et al., 2014)). *Scoring rule $\tilde{\Psi}(\tilde{\mathbf{y}}, \mathbf{q})$ is (strictly) proper to estimate $\mathbf{p}$ from $\tilde{\mathbf{y}}$ if and only if the equivalent loss*

$$\Psi(\mathbf{y}, \mathbf{q}) = \mathbf{y}^T \mathbf{M}^T \tilde{\Psi}(\mathbf{q}), \tag{2.107}$$

is (strictly) proper (See proff by Jesus Cid-Sueiro et al. (2014)). Where $\tilde{\boldsymbol{\Psi}}(\mathbf{q})$ is a vector with components $\tilde{\Psi}_i(\mathbf{q}) = \tilde{\Psi}(\tilde{\mathbf{y}}_i, \mathbf{q})$ and $\tilde{\mathbf{y}}_i$ is the i-th element in $\tilde{\mathrm{Y}}$.

- We can construct a proper loss for weak labels with any pseudoinverse of the mixing matrix, and using its columns as virtual labels (selection via vector multiplication)

$$\tilde{\boldsymbol{\Psi}}(\tilde{\mathbf{y}}_i, \mathbf{q}) = \tilde{\mathbf{v}}_i^\top \boldsymbol{\Psi}(\mathbf{q}),$$

$$\tilde{\boldsymbol{\Psi}}(\tilde{\mathbf{y}}, \mathbf{q}) = (\tilde{\mathbf{y}}^\top \tilde{\mathbf{V}}^\top) \boldsymbol{\Psi}(\mathbf{q}).$$

- Some losses may require a modification in order to ensure the convexity of the weak loss, lowerboundeness and better estimation from a limited set of weak labels (Bacaicoa-Barber et al, 2021)

Losses for weak labels: with unknown weakening process

- **Losses for quasi independent labels:** When the weak label always contains the true label, but with certain probability other labels may appear, then the following virtual label provides CC, RC, and (strictly) proper losses

$$\tilde{\mathbf{v}}_i = \begin{cases} 1, & \text{if } \tilde{\mathbf{y}}_i = 1 \\ -\frac{|\tilde{\mathbf{y}}|-1}{K-|\tilde{\mathbf{y}}|}, & \tilde{\mathbf{y}}_i = 0. \end{cases}$$

- **CC losses for independent labels:** If the weakening process is of the form

$$P(\tilde{\mathbf{y}}|\mathbf{y} = \mathbf{e}_i^K) = \alpha^{\tilde{\mathbf{y}}_i}(1-\alpha)^{1-\tilde{\mathbf{y}}_i}\beta^{|\tilde{\mathbf{y}}|-1}(1-\beta)^{K-|\tilde{\mathbf{y}}|}$$

  It is possible to use the weak labels directly as virtual labels to obtain CC losses (considering non-degenerate cases).

Empirical analysis with weak labels

- **Perello-Nieto et. al. 2017** shows empirical results with known and unknown weakening processes
- **Bacaicoa et. al. 2021** shows empirical results with known weakening processes
- **Perello-Nieto et. al. 2020** shows how to combine multiple sources of weak labels in one dataset

Other approaches to learn in the proposed setting

- The large ratio of annotated vs unannotated data could be exploited with semi-supervised methods
- We have tested active learning methods to select candidate samples to be annotated in **Bi et. al. 2020**

More details about weak labels:

- **Poyazki et. al. 2022** describes a landscape of weak labels

Deployment

Model predictions in the wild

- Class and ranking calibrated predictions:
  - Can be optimal if trained in the same conditions as deployment
  - Can not be adjusted to new contexts
- Scoring and probability calibrated predictions:
  - Can be adapted to new operating conditions
  - Can abstain to avoid ambiguous predictions
- Probability estimators with uncertainty
  - May detect data shift
  - May detect new classes or unknown patterns
  - Can abstain to avoid errors because of lack of knowledge

Probability estimation

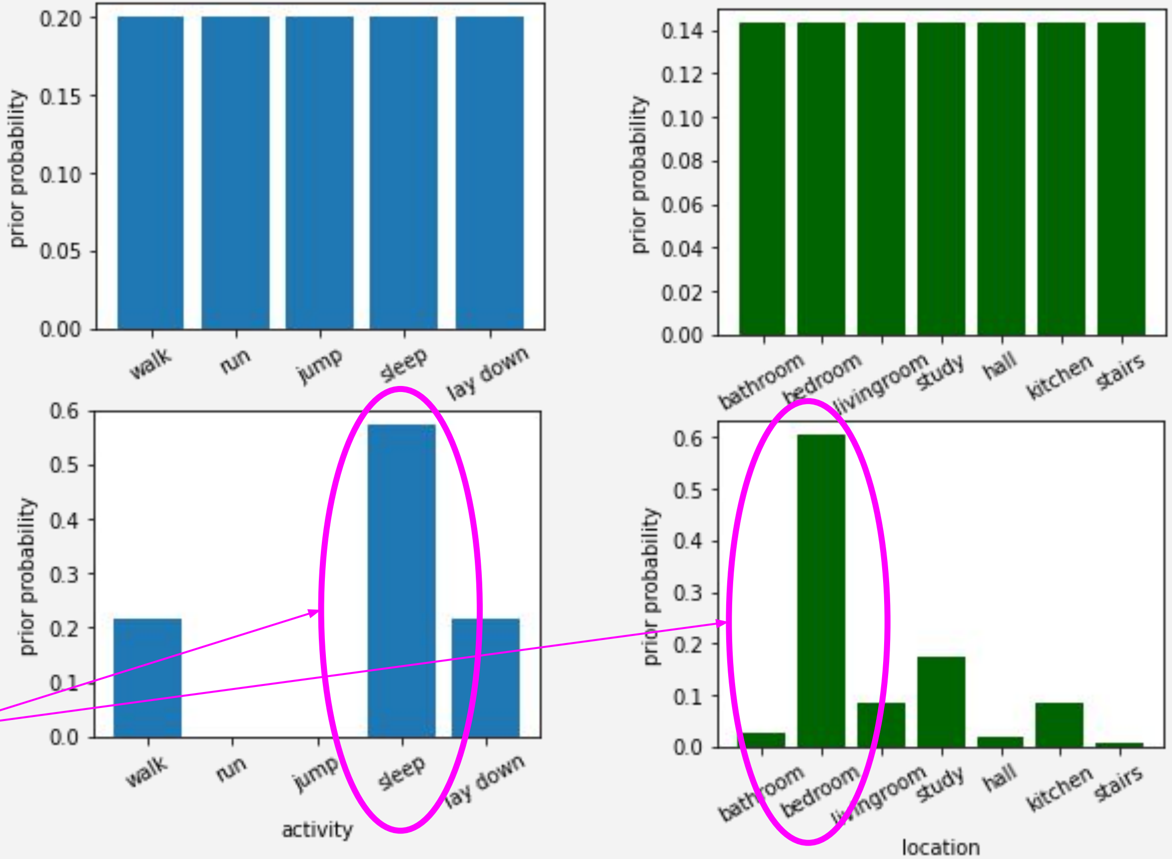
Change of priors

Activites

Indoor locations

Training

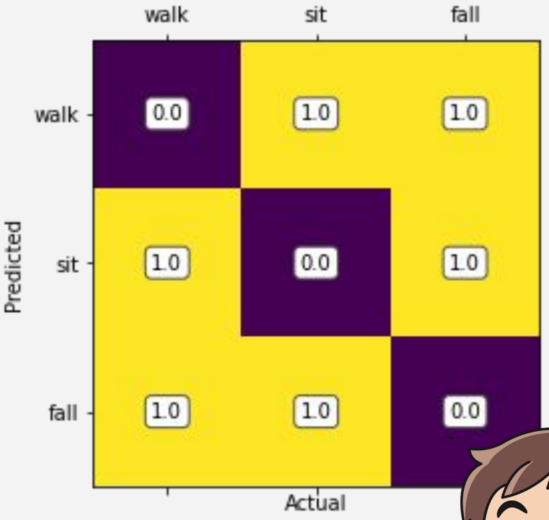
Deployment



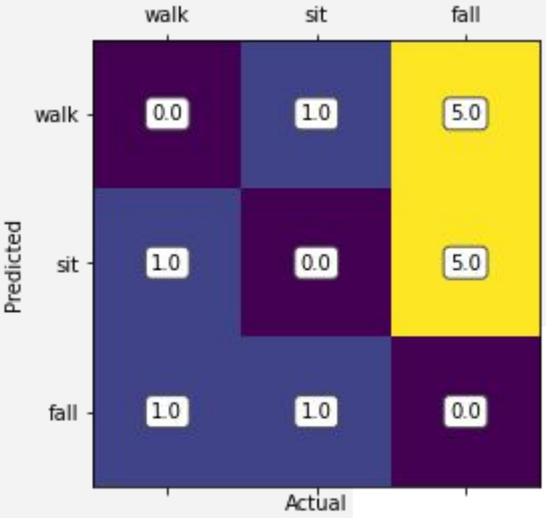
About 60% accuracy with a constant classifier!

Change of costs: e.g. cost of falling

Training (0-1 loss)



Deployment

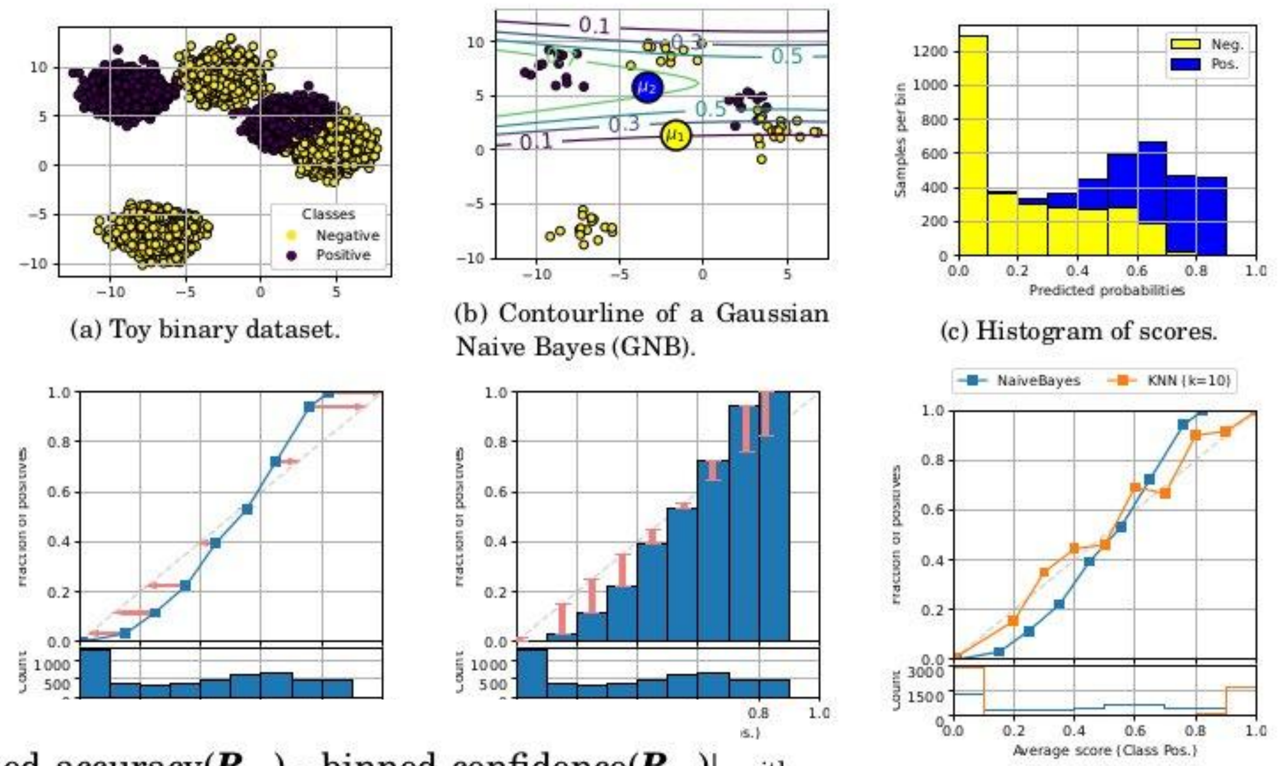


https://creazilla.com/nodes/8121-successful-young-man-clipart

https://designbundles.net/plusstore/1128762-old-man-clipart

Evaluating probability correctness

- Binning the model scores
- Reliability diagram
- Necessary correction
- Error gaps
- Comparison of two models
- Metrics: confidence Expected Calibration Error (conf-ECE) and its maximum (conf-MCE)



(a) Toy binary dataset.

(b) Contourline of a Gaussian Naive Bayes (GNB).

(c) Histogram of scores.

(f) Comparison of the calibration of two classifiers.

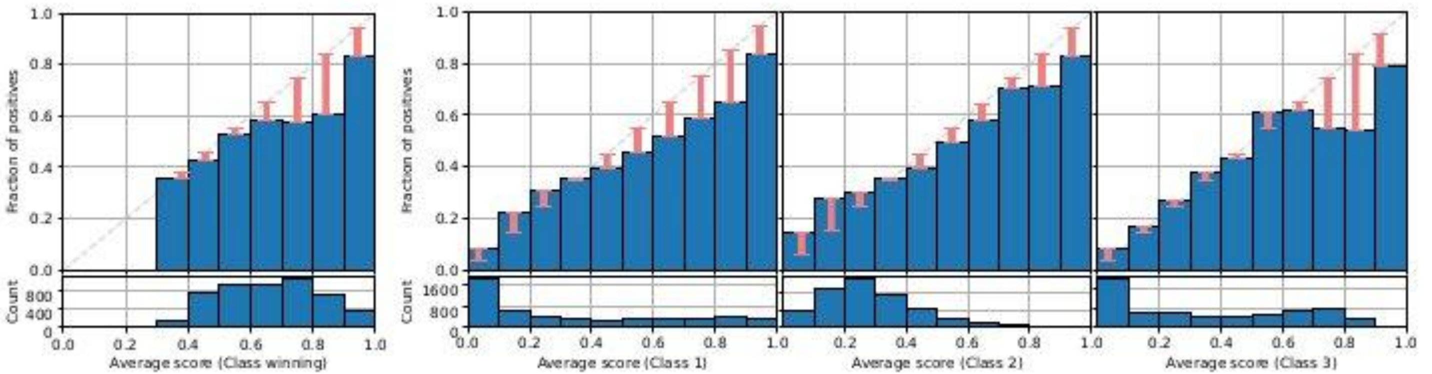
$$\text{confidence-ECE}(\mathscr{B}) = \sum_{m=1}^{M} \frac{|\boldsymbol{B}_m|}{N} \left|\text{binned-accuracy}(\boldsymbol{B}_m) - \text{binned-confidence}(\boldsymbol{B}_m)\right| .$$ with er-

Evaluating multiclass probabilities

- Metrics: classwise ECE and MCE

$$\text{class-}j\text{-ECE}(\mathscr{B}) = \sum_{i=1}^{M} \frac{|\boldsymbol{B}_{i,j}|}{N} |\bar{y}_j(\boldsymbol{B}_{i,j}) - \bar{p}_j(\boldsymbol{B}_{i,j})|,$$

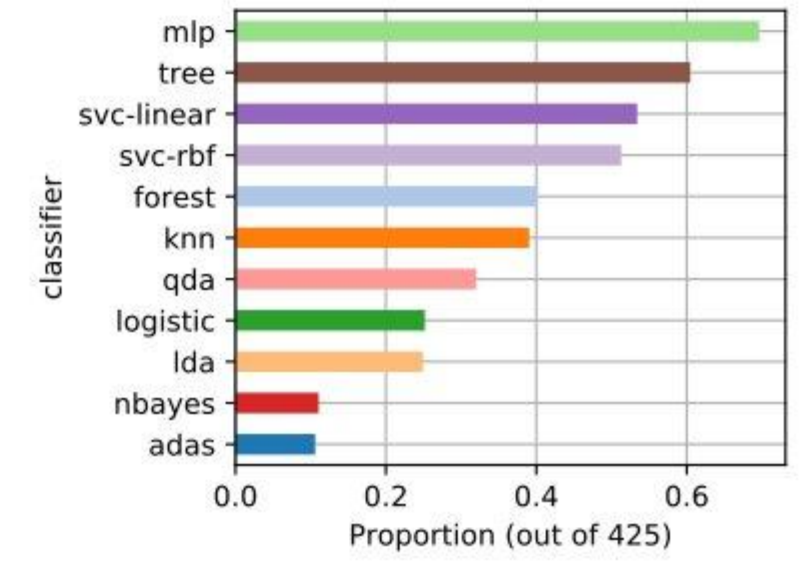
$$\text{classwise-ECE}(\mathscr{B}) = \frac{1}{K} \sum_{j=1}^{K} \text{class-}j\text{-ECE}$$



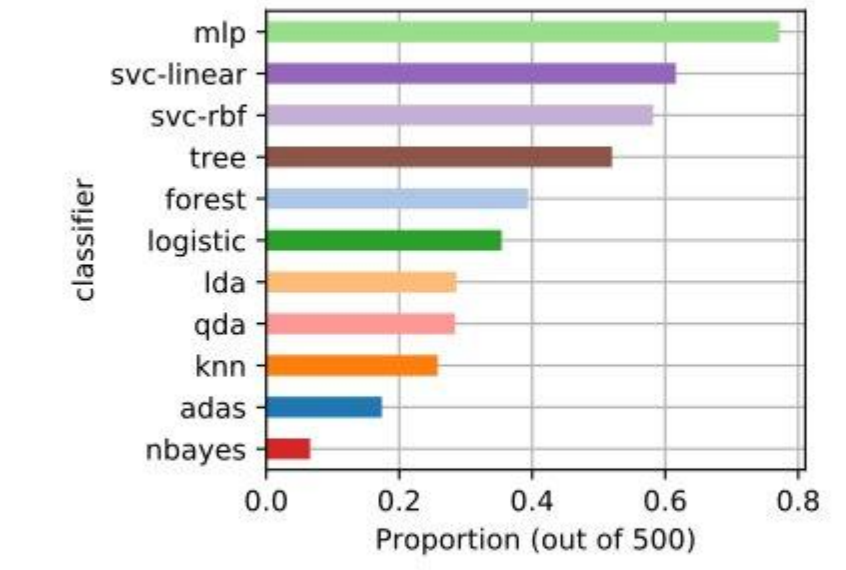
(a) Confidence reliability diagram.

(b) One-vs-rest reliability diagram of a GNB with calibration gaps.

How calibrated are common probabilistic classifiers



(a) p-conf-ece



(b) p-cw-ece

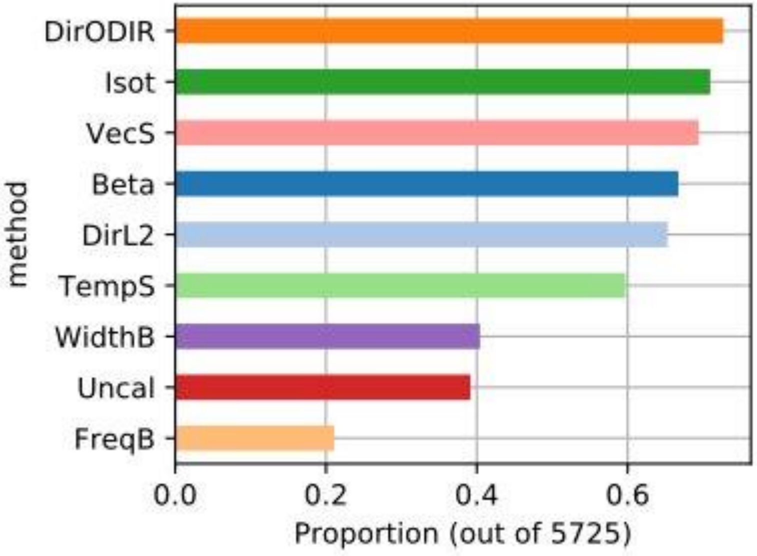
Existing multiclass calibration methods

- Methods to improve probabilities from decision trees
- Binning calibration methods with one-vs-the-rest aggregation (OvR)
- Platts scaling on infinite support scores (multinomial logistic regression)
- Isotonic regression with OvR
- Beta calibration with OvR
- Temperature, vector and matrix scaling for DNNs
- Dirichlet calibration (Kull et. al. 2019)
    - Assume a Dirichlet distribution per class $p(\mathbf{q}|C_i) \sim \mathrm{Dir}(\alpha_k)$
    - Generative learning assumption

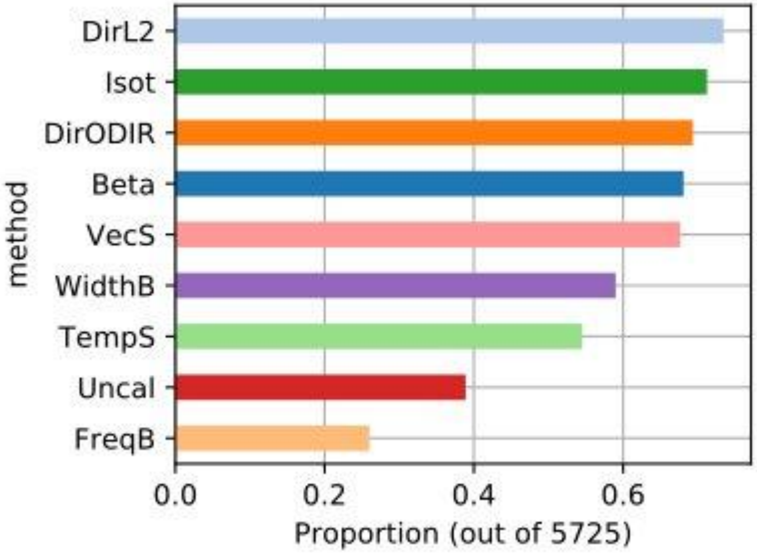
$$\mu_{\mathrm{DirGen}}(\mathbf{q};\mathbf{A},\pi) = \left(\frac{\mathrm{dir}(\mathbf{q};\boldsymbol{\alpha}_1)\pi_1}{p(\mathbf{q})}, \ldots, \frac{\mathrm{dir}(\mathbf{q};\boldsymbol{\alpha}_K)\pi_K}{p(\mathbf{q})}\right),$$

    - Check Kull et. al. 2019 for cannonical and linear parameterizations

Comparison of multiclass calibrators



(a) p-conf-ece



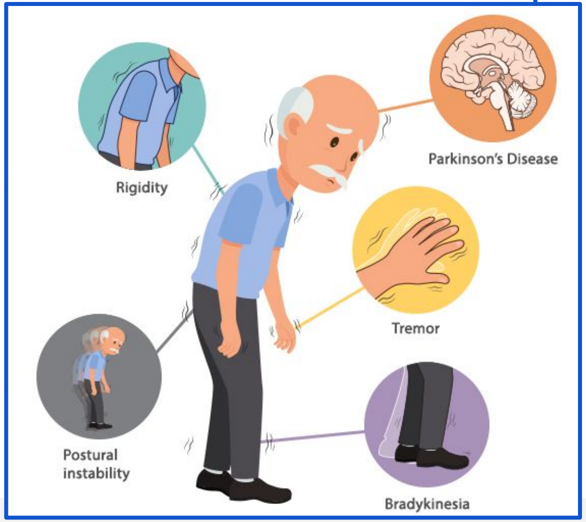
(b) p-cw-ece

Uncertainty estimation

Unknown activities/patterns

- **New classes** may appear during deployment
- **New participants** may be different to the participants used during the labelling
  - E.g. young person standing still or a person with Parkinson's disease
- May be interested to detect classes that are different to our training data

https://www.vecteezy.com/vector-art/2186373-set-of-daily-routines-the-concept-of-daily-life-everyday-leisure-and-work-activities-flat-vector-illustration

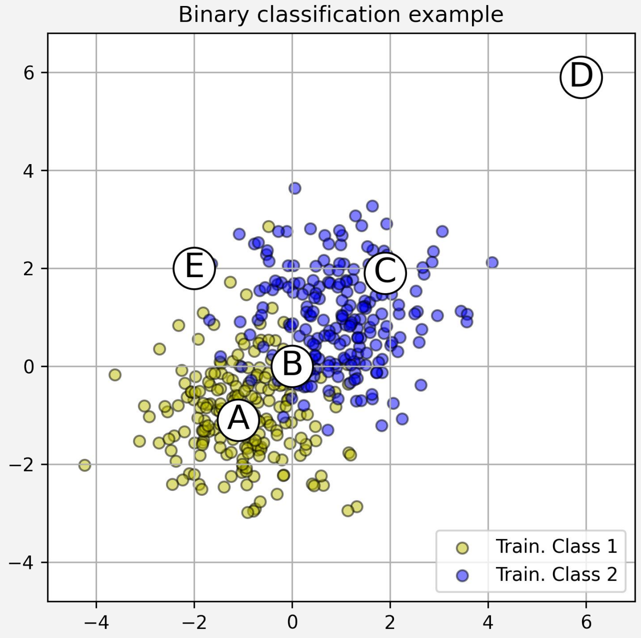


https://neurologysleepcentre.com/blog/what-is-parkinsons-disease/

Binary classification example

Binary classification problem with two features, but generalises to arbitrary number of classes and dimensions.

- A, B, and C are in **dense** regions
- E and D are in **low density** regions
- B and E are in the **decision boundary**

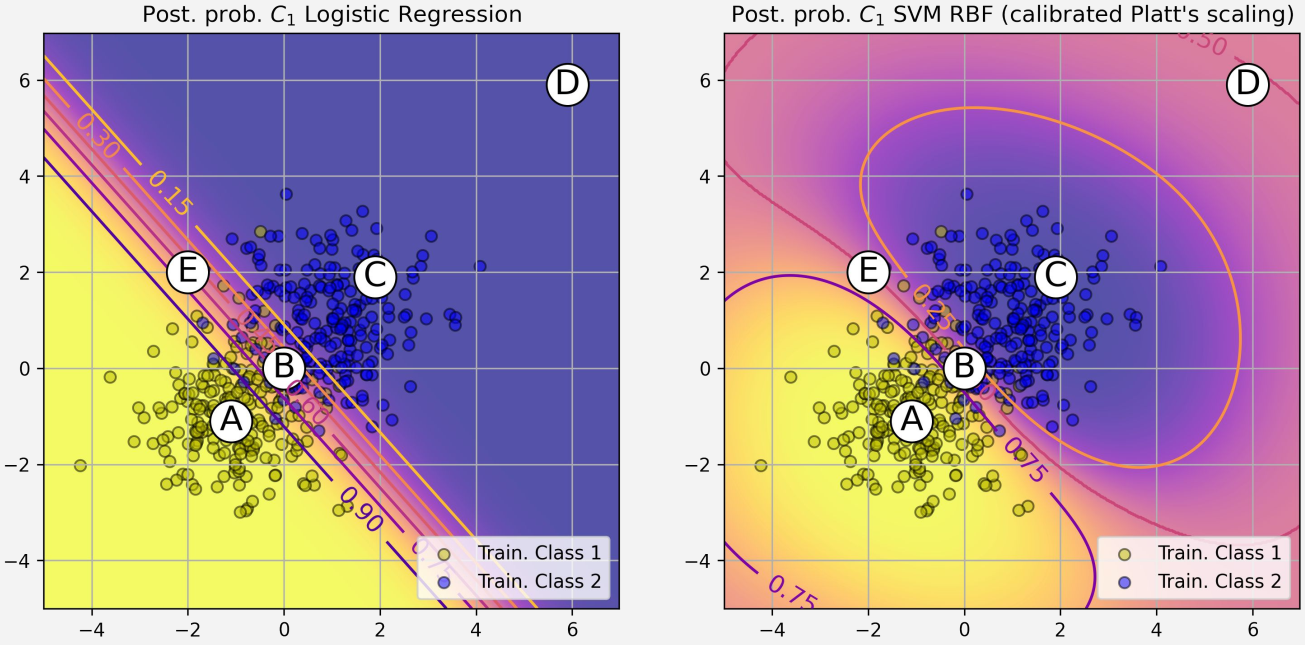


Two common classifiers

Minimization of empirical risk.

Focus on the performance in regions of high density.

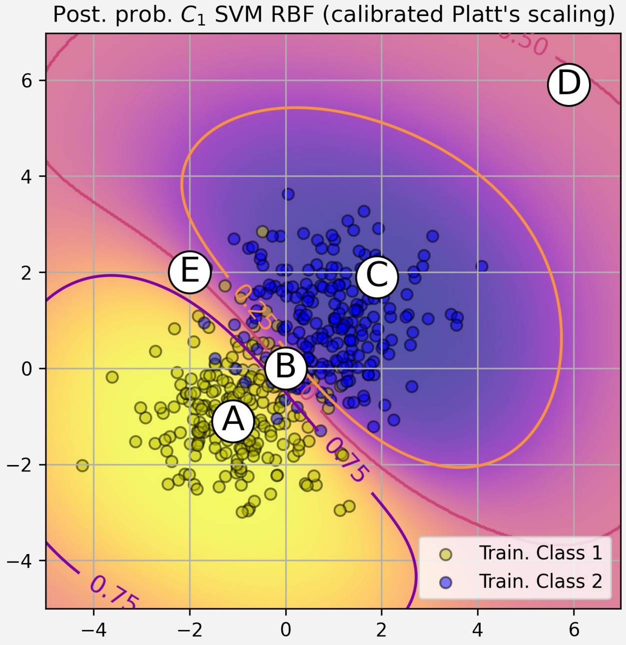
Expect same data distribution during deployment.



Interpretation of the posterior probabilities

- A is clearly from Class 1
- C is clearly from Class 2
- B, E and D are in the same issoline 0.5
- **Several examples in B**
- **No examples in D**

	$p(C_1\|x)$	$p(C_2\|x)$
A	1	.0
B	*.5*	*.5*
C	.0	1
D	*.5*	*.5*
E	*.5*	*.5*

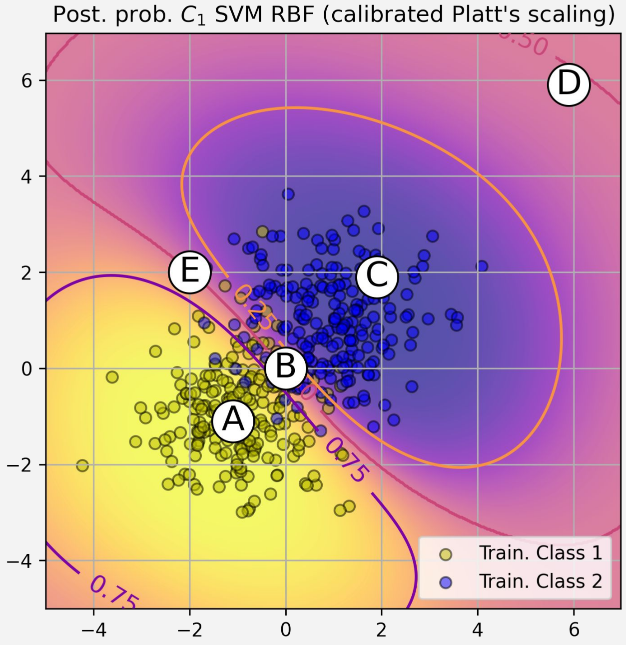


Adding an additional posterior probability (background)

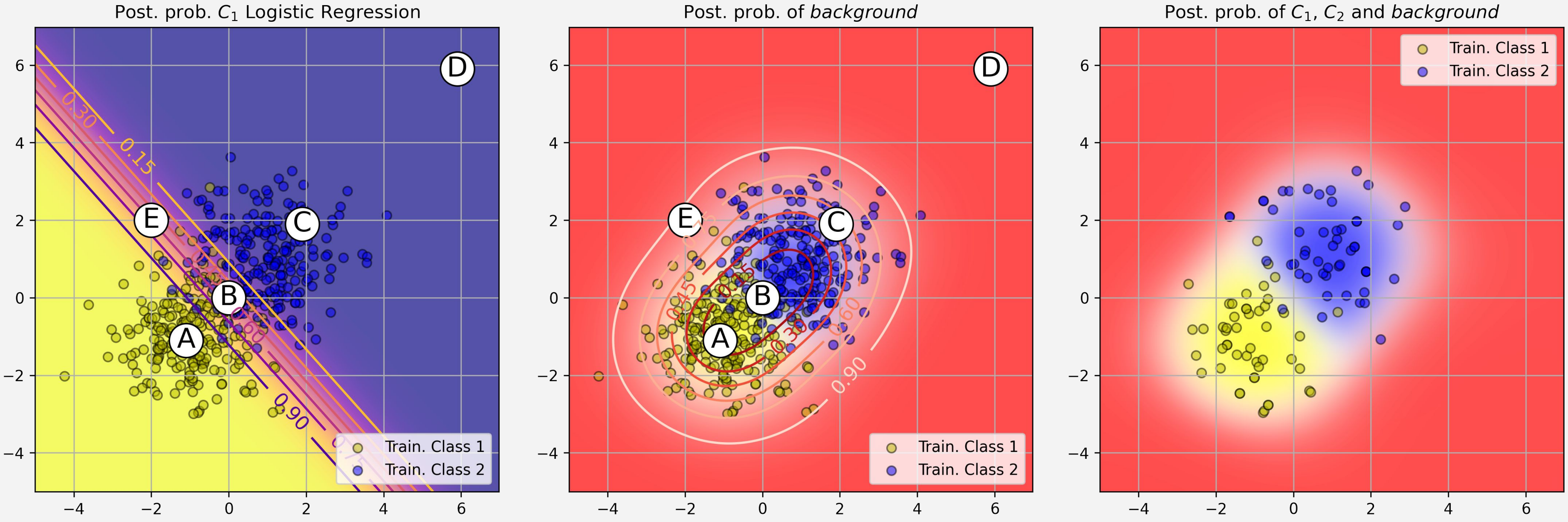
We refer to the foreground class as the known training data, and background class the rest.

- We are certain about B being ambiguous
- We are uncertain about D

	$p(C_1\|x)$	$p(C_2\|x)$	$p(b\|x)$
A	$1 \to .9$	$.0 \to .0$	$.1$
B	$.5 \to .5$	$.5 \to .5$	$.0$
C	$.0 \to .0$	$1 \to .5$	$.5$
D	$.5 \to .0$	$.5 \to .0$	1
E	$.5 \to .1$	$.5 \to .1$	$.8$



Objective: Adapt an arbitrary classifier to provide familiarity



How to adapt the probabilities in theory

Base classifier: known posterior class probabilities

$$p(f_c|f,x) = \frac{p(x|f,f_c)p(f_c|f)}{p(x|f)} \quad \text{for} \quad c = 1,\ldots,C$$

We want: foreground vs background posterior probabilities

$$p(f|x) = \frac{p(x|f)p(f)}{p(x)} \qquad p(b|x) = \frac{p(x|b)p(b)}{p(x)}.$$

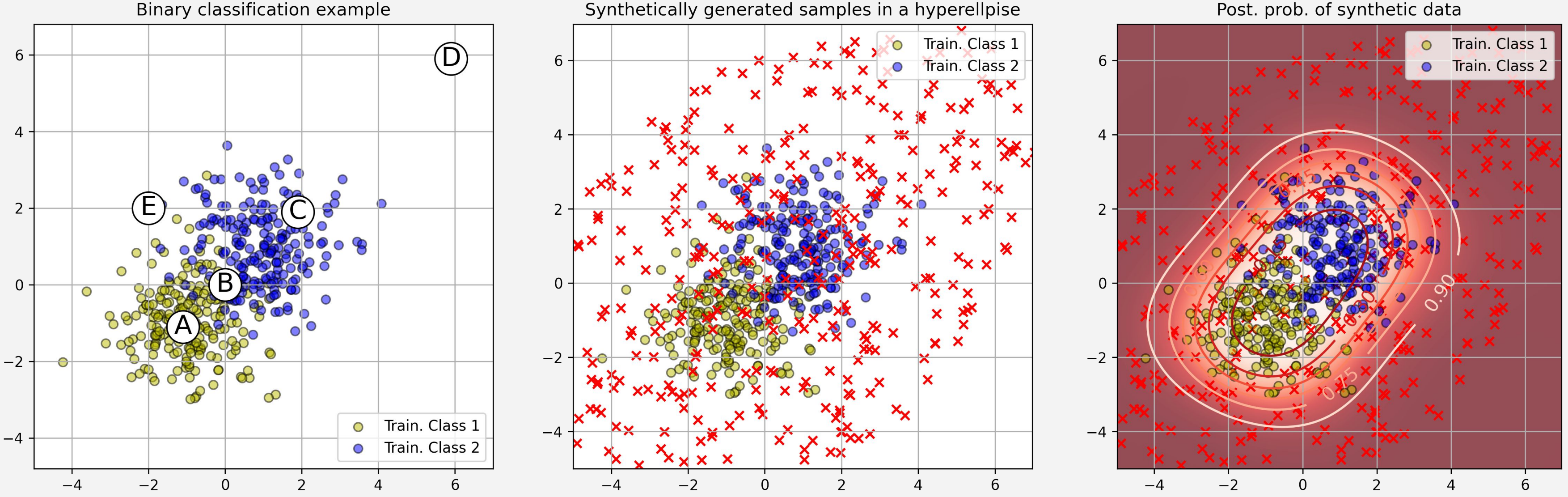
We only need the ratio between the previous probabilities

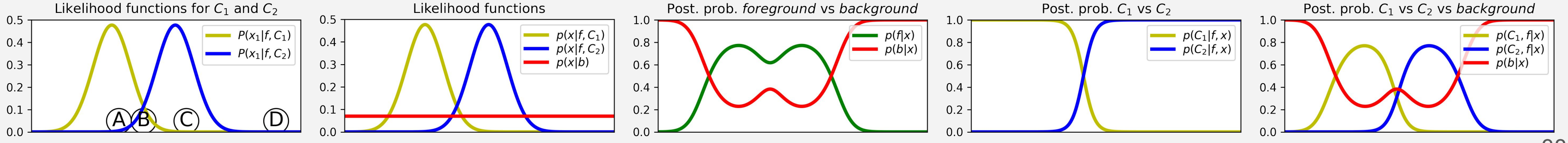
$$r(x) = p(f|x)/p(b|x)$$

We obtain posteriors for all foreground classes and background class

$$p(f_c|x) = \frac{p(f_c|f,x)r(x)}{1+r(x)} \qquad p(b|x) = \frac{1}{1+r(x)}$$

A discriminative approach and synthetic data

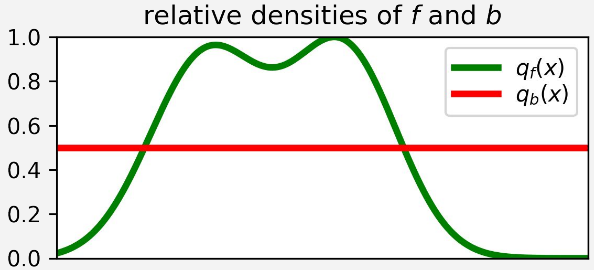




A familiarity approach and density estimation

- Estimate density of foreground (training data)
- Obtain **relative density** with respect to the maximum of foreground

$$q_f(x) = \frac{p(f,x)}{\max_x p(x,f)},$$
$$q_b(x) = \frac{p(b,x)}{\max_x p(x,f)}$$



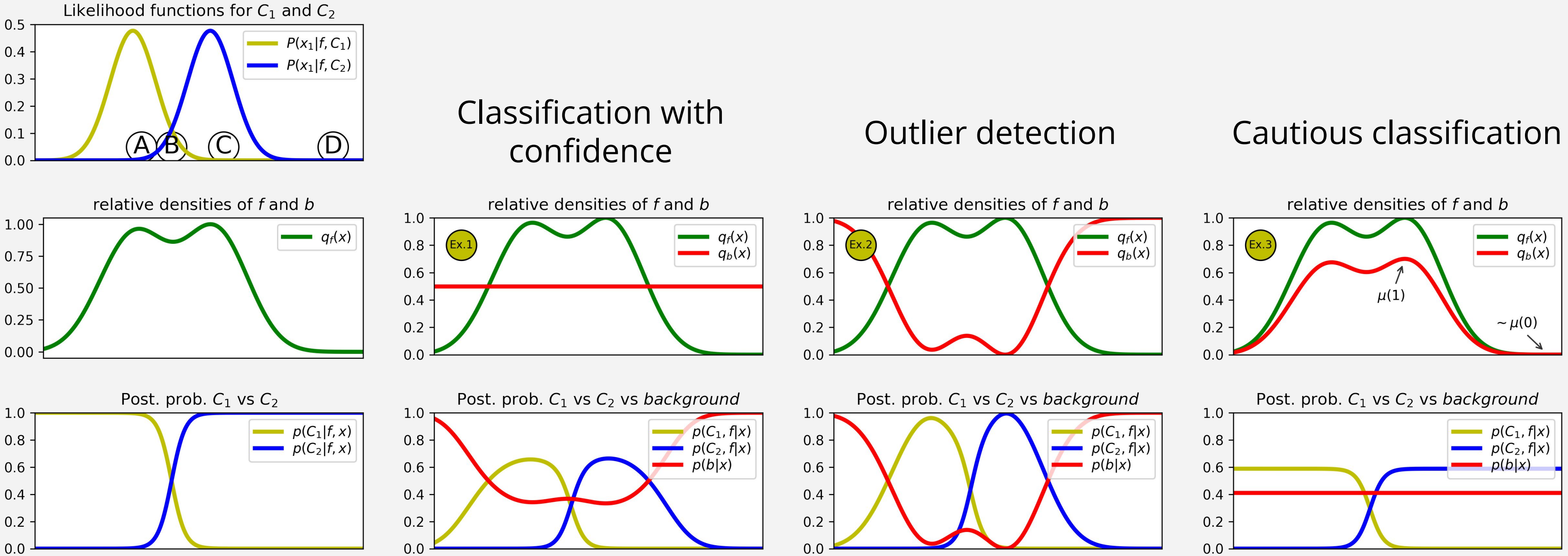
With those, we can still obtain the familiarity ratio

$$\cancel{r(x) = p(f|x)/p(b|x)} \qquad r(x) = q_f(x)/q_b(x)$$

Obtain the new posterior probabilities

$$p(f_c|x) = \frac{p(f_c|f,x)r(x)}{1+r(x)} \qquad p(b|x) = \frac{1}{1+r(x)}$$

A familiarity approach



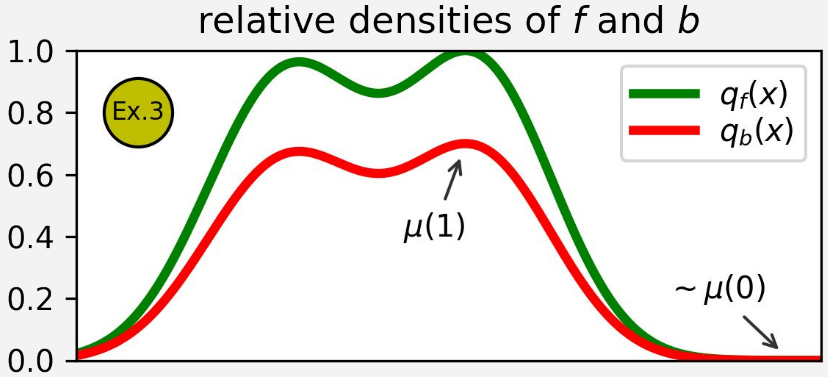
A familiarity approach and affine transformation

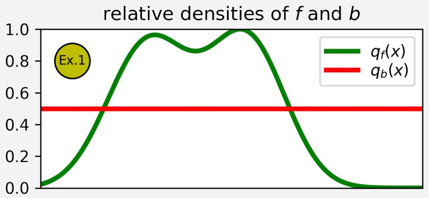
Relative density of the background as a function of the foreground

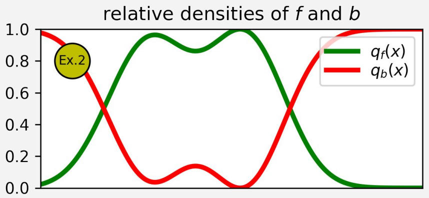
$$q_b(x) = \mu(q_f(x))$$

Parametric form with minimum and maximum values.

$$\mu(z) = (1 - z)\mu(0) + z\mu(1)$$

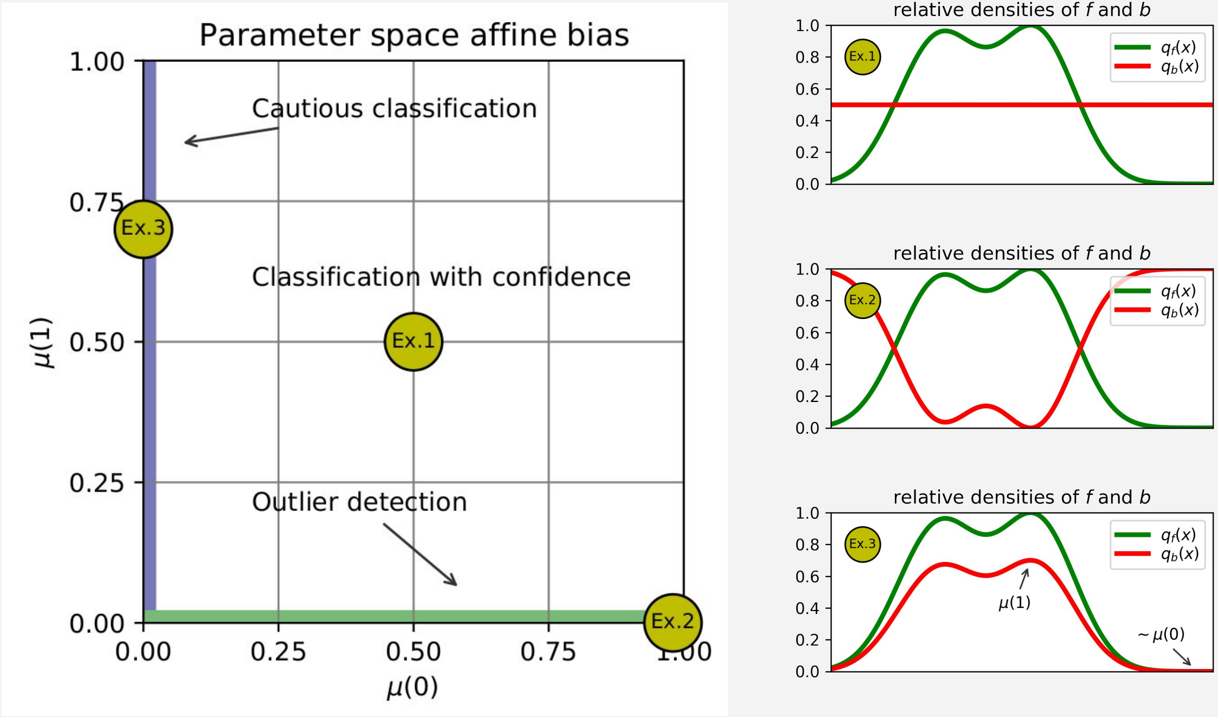






A familiarity approach and affine transformation

Other possible values



$$q_b(x) = \mu(q_f(x))$$

$$\mu(z) = (1 - z)\mu(0) + z\mu(1)$$

Results

Our tests with 41 multiclass datasets showed:

1. Significantly better performance in **classification with confidence** against a SOTA method
2. Competitive results for **outlier detection** against two specialised methods

And it is equivalent to Chow's rule to perform **cautious classification**

More details in:

M. Perello-Nieto, T. M. S. Filho, M. Kull and P. Flach, "**Background Check: A General Technique to Build More Reliable and Versatile Classifiers,**" 2016 IEEE 16th International Conference on Data Mining (ICDM), 2016, pp. 1143-1148, doi: 10.1109/ICDM.2016.0150.

reframe.github.io/background_check

Conclusion

1. Consider the **model assumptions** in real-world problems
2. The available data for **training may be biased**
3. The annotation process may generate labels of different quality (**weak labels**)
4. **Probabilities** allow an easy adaptation with operating condition changes
5. **Abstaining** can be necessary in critical decision making
6. Quantify the **uncertainty** in the predictions

References

Perello-Nieto, Miquel, Raul Santos-Rodriguez, Dario Garcia-Garcia, and Jesus Cid-Sueiro. 2020. “Recycling Weak Labels for Multiclass Classification.” Neurocomputing 400:206–15. doi: 10.1016/j.neucom.2020.03.002.

Poyiadzi, Rafael, Daniel Bacaicoa-Barber, Miquel Perello-Nieto, Raul Santos-Rodriguez, Jesus Cid-Sueiro, and Peter Flach. 2022. “The Weak Supervision Landscape.” P. NA in 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops).

Perelló-Nieto, Miquel, Raúl Santos-Rodríguez, and Jesús Cid-Sueiro. 2017. “Adapting Supervised Classification Algorithms to Arbitrary Weak Label Scenarios.” Pp. 247–59 in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 10584 LNCS, edited by N. Adams, A. Tucker, and D. Weston. London, UK: Springer, Cham.

Holmes, Michael, Miquel Perello Nieto, Hao Song, Emma Tonkin, Sabrina Grant, and Peter Flach. 2020. “Modelling Patient Behaviour Using IoT Sensor Data: A Case Study to Evaluate Techniques for Modelling Domestic Behaviour in Recovery from Total Hip Replacement Surgery.” Journal of Healthcare Informatics Research 4(3):238–60. doi: 10.1007/s41666-020-00072-6.

Bi, Haixia, Miquel Perello-Nieto, Raul Santos-Rodriguez, and Peter Flach. 2021. “Human Activity Recognition Based on Dynamic Active Learning.” IEEE Journal of Biomedical and Health Informatics 25(4):922–34. doi: 10.1109/jbhi.2020.3013403.

Perello-Nieto, Miquel, Telmo M. Silv. Silva Filho, Meelis Kull, and Peter Flach. 2016. “Background Check: A General Technique to Build More Reliable and Versatile Classifiers.” Pp. 1143–48 in 2016 IEEE 16th International Conference on Data Mining (ICDM 2016), Proceedings of the IEEE International Conference on Data Mining (ICDM). United States: Institute of Electrical and Electronics Engineers (IEEE).

Bacaicoa-Barber, Daniel, Miquel Perello-Nieto, Raul Santos-Rodriguez, and Jesus Cid-Sueiro. 2021. “On the Selection of Loss Functions under Known Weak Label Models.” in 30th International Conference on Artificial Neural Networks, ICANN2021.

Cid-Sueiro, Jesús, Darío García-García, and Raúl Santos-Rodríguez. 2014. “Consistency of Losses for Learning from Weak Labels.” Pp. 197–210 in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 8724 LNAI. Springer Berlin Heidelberg.

References

Twomey, Niall, Hao Song, Massimo Camplani, Sion L. Hannuna, Ni Zhu, Pete Woznowski, Miquel Perello-Nieto, Emma L. Tonkin, Peter A. Flach, and Ian J. Craddock. 2020. “SPHERE House Scripted Dataset: A Multi-Sensor Dataset with Annotated Activities of Daily Living Recorded in a Residential Setting.”

Tonkin, Emma, Antonis Vafeas, Miquel Perello-Nieto, Haixia Bi, and Ian Craddock. 2020. “Sphere House Multi-Wearable.”

Kull, Meelis, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. “Beyond Temperature Scaling: Obtaining Well-Calibrated Multiclass Probabilities with Dirichlet Calibration.” Pp. 12316–26 in Advances in Neural Information Processing Systems, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett. Vancouver, British Columbia, Canada: Curran Associates, Inc.

Flach, Peter, Miquel Perello-Nieto, Hao Song, Meelis Kull, and Telmo Silva-Filho. 2020. “Classifier Calibration: How to Assess and Improve Classifier Confidence and Uncertainty.” in Tutorial at The European Conference on Machine Learning and Principles and Practice of Know-ledge Discovery in Databases (ECML-PKDD). Ghent, Belgium.

Tonkin, Emma L., Miquel Perello Nieto, Haixia Bi, and Antonis Vafeas. 2020. “Towards a Methodology for Acceptance Testing and Validation of Monitoring Bodyworn Devices.” Pp. 1–6 in 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops).

Holmes, Mike, Hao Song, Emma Tonkin, Miquel Perello-Nieto, Sabrina Grant, and Peter Flach. 2018. “Analysis of Patient Domestic Activity in Recovery from Hip or Knee Replacement Surgery: Modelling Wrist-Worn Wearable RSSI and Accelerometer Data in the Wild.” in Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data (KDH at IJCAI-ECAI 2018). Vol. 2148, edited by K. Bach, R. Bunescu, O. Farri, A. Guo, S. Hasan, Z. M. Ibrahim, C. Marling, J. Raffa, J. Rubin, and H. Wu. Stockholm, Schweden.

Morgan, Catherine, Farnoosh Heidarivincheh, Ian Craddock, Ryan McConville, Miquel Perello Nieto, Emma L. Tonkin, Alessandro Masullo, Antonis Vafeas, Mickey Kim, Roisin McNaney, Gregory J. L. Tourte, and Alan Whone. 2021. “Data Labelling in the Wild: Annotating Free-Living Activities and Parkinson’s Disease Symptoms.” Pp. 471–74 in 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops).