

A Supplementary Material

A.1 Fitting a distribution to distances

The definition of farness (6) requires an estimated cumulative distribution function of the distance $D(\mathbf{y}, g)$ where \mathbf{y} is a random object generated from class g . The available data are the $D(i, g_i)$ of each object i to its class label g_i . In view of possible heteroskedasticity between classes, we start by normalizing per class. For a given class g we divide all the $D(i, g)$ where i is a member of class g by $\text{median}\{D(j, g); j \text{ belongs to class } g\}$. The resulting distances are more homoskedastic, and we pool them to obtain distances d_i for $i = 1, \dots, n$. The empirical distribution of the d_i is typically right-skewed.

In order to account for skewness, we apply the function `transfo` of the R-package `cellWise` (Raymaekers and Rousseeuw, 2020) with default options. This function first standardizes the d_i to

$$x_i = \frac{d_i - \text{Med}}{\text{Mad}}$$

where $\text{Med} = \text{median}_{j=1}^n d_i$ and Mad is the median absolute deviation given by $\text{Mad} = 1.4826 \text{median}_{j=1}^n |d_i - \text{Med}|$ as implemented in the standard function `mad()` in R. Next, `transfo` carries out the Yeo-Johnson transform given by

$$h_\lambda(x) = \begin{cases} ((1+x)^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \text{ and } x \geq 0 \\ \log(1+x) & \text{if } \lambda = 0 \text{ and } x \geq 0 \\ -((1-x)^{2-\lambda} - 1)/(2-\lambda) & \text{if } \lambda \neq 2 \text{ and } x < 0 \\ -\log(1-x) & \text{if } \lambda = 2 \text{ and } x < 0 \end{cases} \quad (\text{A.1})$$

which aims to bring the distribution close to a normal distribution. The transformation h_λ is characterized by a parameter λ that has to be estimated from the data. This estimation is typically done by maximum likelihood, but the default in `transfo` is to apply the weighted maximum likelihood estimator of Raymaekers and Rousseeuw (2021b) which is less sensitive to outliers. The resulting $h_\lambda(x_i)$ are in turn standardized by their own Med and Mad , yielding z_i whose distribution is approximately standard normal. The estimated cdf of the distances d_i is then given by $\hat{F}(d_i) := \Phi(z_i)$ where Φ is the standard normal cdf.

A.2 More on the CIFAR-10 data

Here we show some visualizations of classes in the CIFAR-10 data that were not in the main text.

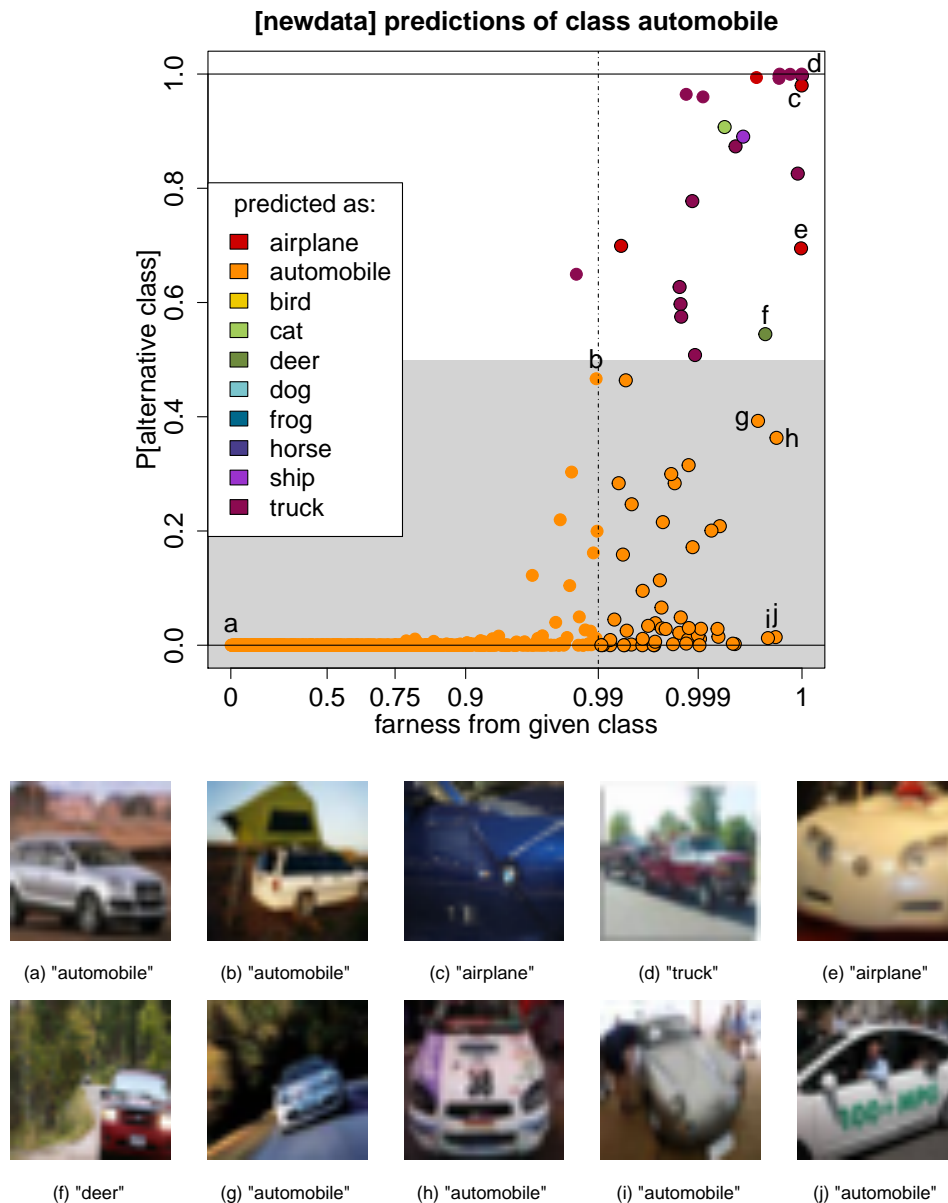


Figure 13: Class map of the automobile class, with the corresponding images.

Note that (d) in Figure 13 and (b) and (c) in Figure 14 look like pickup trucks, which are in a sense intermediate between automobiles and trucks, in spite of the fact that the original data description in <https://www.cs.toronto.edu/~kriz/cifar.html> aimed to avoid pickup trucks for that reason.

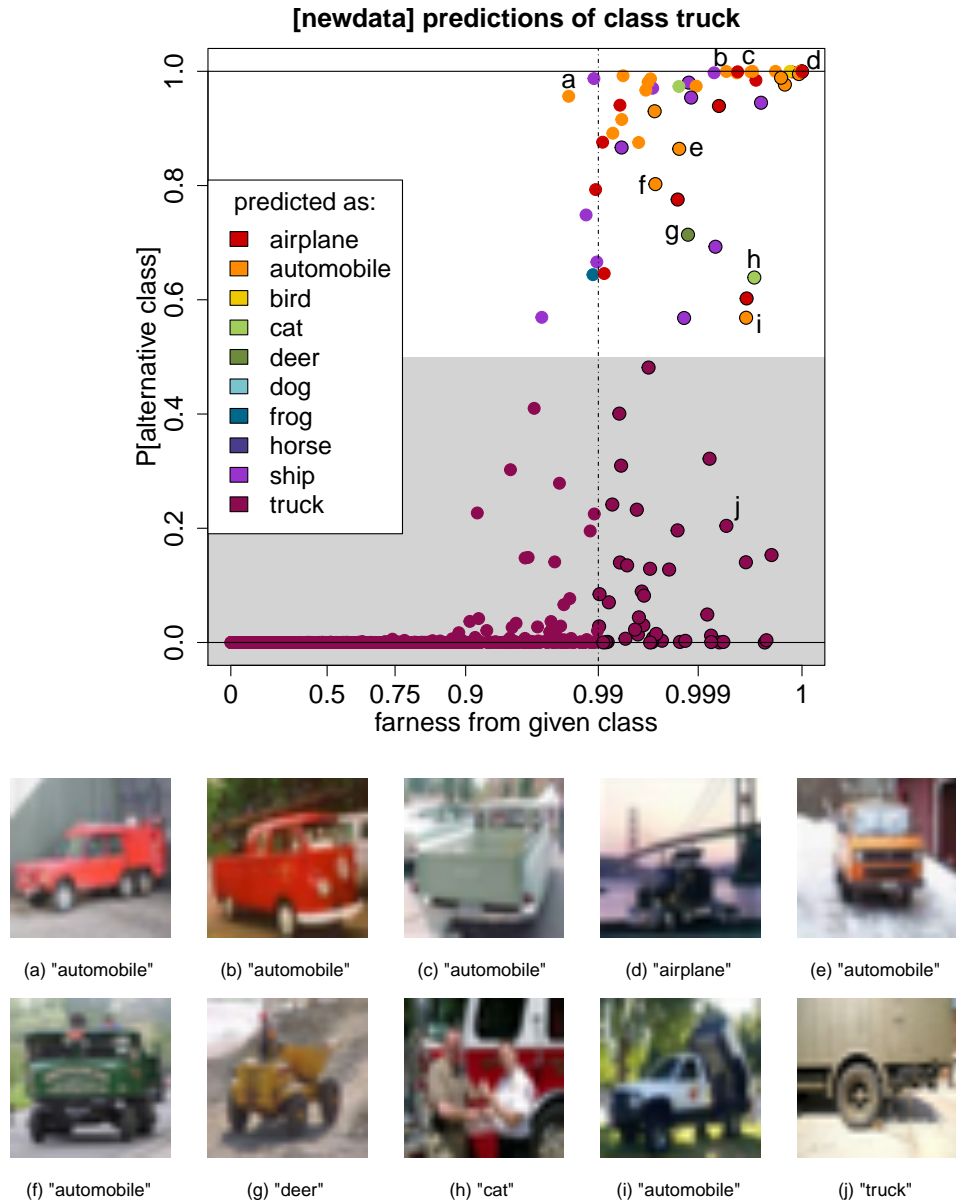


Figure 14: Class map of the truck class, with the corresponding images.

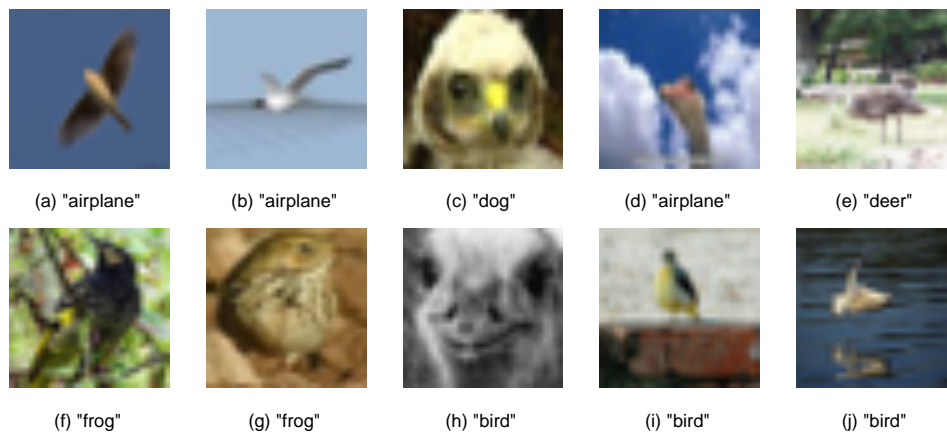
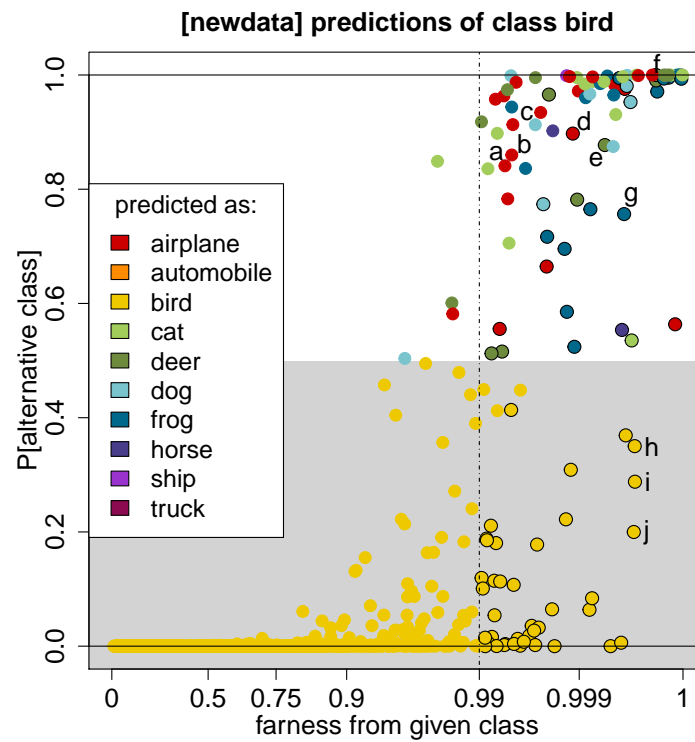


Figure 15: Class map of the bird class, with the corresponding images.

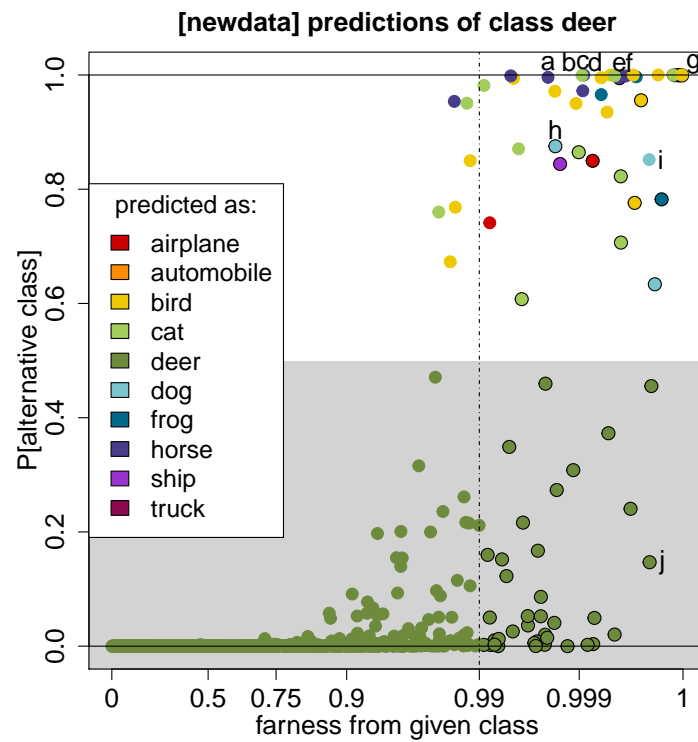


Figure 16: Class map of the deer class, with the corresponding images.

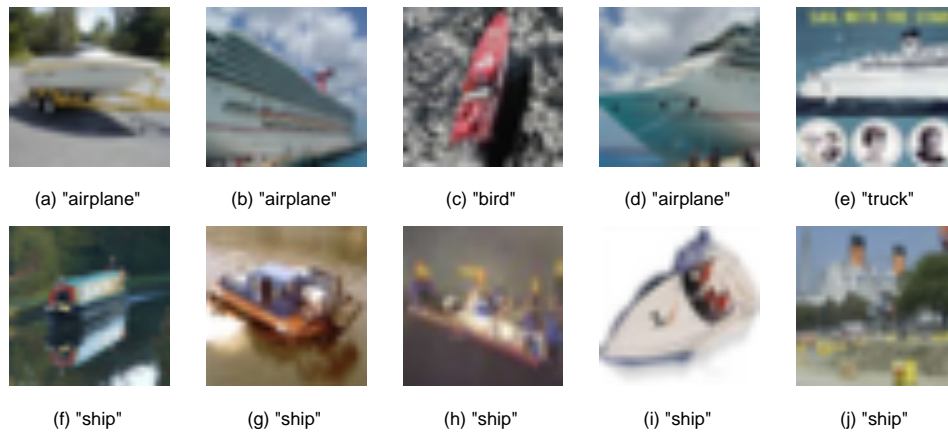
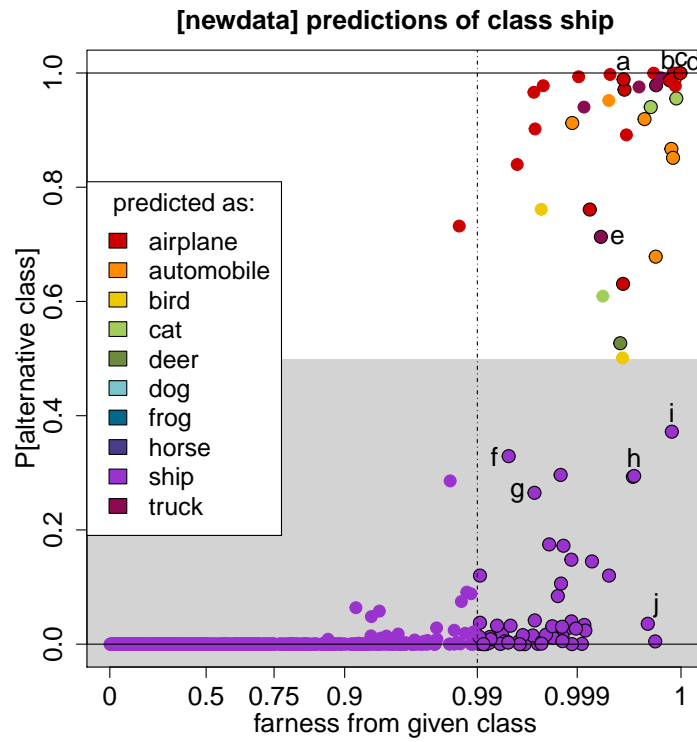


Figure 17: Class map of the ship class, with the corresponding images.

A.3 Computing farness for tree-based classifiers

In subsections 4.2 and 4.3 we described how to compute interpoint dissimilarities $d(i, j)$ between any two cases i and j in the training data, by applying **daisy** with variable weights equal to each variable’s importance as obtained from the classifier.

The task at hand is to derive a dissimilarity measure $D(i, g)$ of each training case i to every class g . Given that the classes may form disconnected regions in feature space, the construction needs to be local rather than global. For each object i and class g we compute $D(i, g)$ as the median of the k smallest dissimilarities $d(i, j)$ to all objects j of class g . The number k can be chosen by the user. The default is $k = 5$, which worked well in a wide range of applications. For each class g we then divide $D(i, g)$ by $\text{median}\{D(j, g); j \text{ belongs to class } g\}$. This makes the $D(\cdot, g)$ values from all classes more comparable to each other. Finally, we estimate the distribution of the $D(i, g)$ as in the previous section A.1, yielding $\text{farness}(i, g)$.

The above formulas can also be used for new data, such as a test set. We then start by computing all dissimilarities $d(i, h)$ where case i belongs to the new dataset and h is any case in the training data. This computation uses the same variable weights and other parameters as in the training data. We then compute $D(i, g)$ as the median of the k smallest dissimilarities $d(i, h)$ to all objects h of class g in the training data. Here k is the same as in the training data. We then divide $D(i, g)$ by the same denominator $\text{median}\{D(j, g); j \text{ belongs to class } g\}$ that was already computed on the training data. In order to turn the $D(i, g)$ into $\text{farness}(i, g)$ we apply the transformation fitted to the training data in section A.1, that is, we standardize the d_i with the median and mad from the training data, then apply the Yeo-Johnson transform (A.1) with the same λ , and then standardize the result with the same constants as in the training data.

All of this ensures that the farness of a new case in the test set only depends on the training data and the new case, and not on other cases in the test set. In principle, the new dataset could even consist of a single case.

A.4 The Titanic test data

We now analyze the Titanic test data. The classification tree obtained on the training data and shown in Figure 6 has an accuracy of about 78% on the test data, which is not much lower than the 82% on the training data. Figure 18 shows the silhouette plot on the test data. Its overall average silhouette width is slightly lower than on the training data, so the classification is less precise. On the other hand, the shape of the silhouette plot looks like that of the training data, so the classifier behaves in a similar fashion here. The class of survivors again proved harder to predict than the class of casualties.

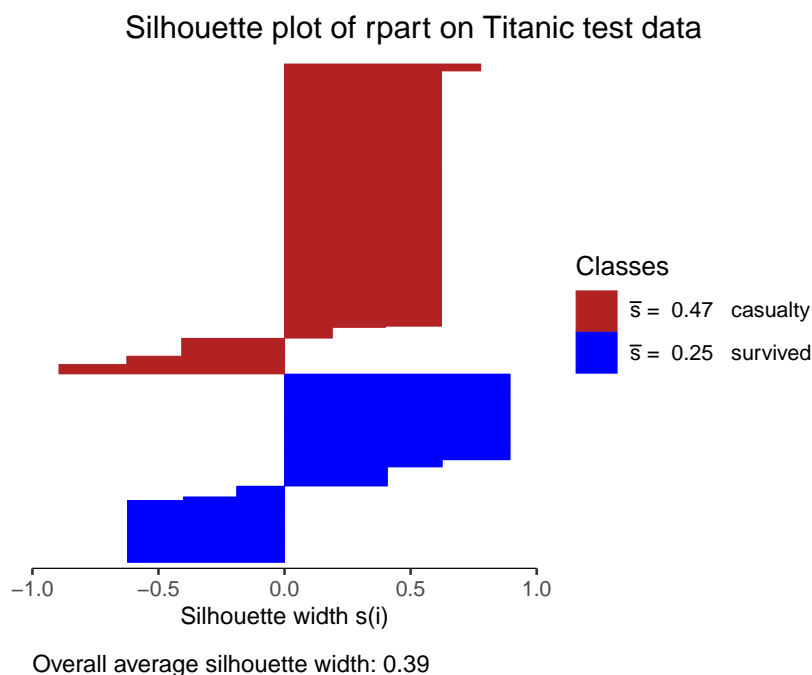


Figure 18: Silhouette plot of the classification of the Titanic test data.

The class maps of both classes are shown in Figure 19. The left panel is from the casualty class. Passenger **a** sits well within the class of casualties and is predicted as casualty with low PAC, i.e. fairly high conviction. It is a male passenger traveling in third class without unusual variables. Case **b** is a female traveling in third class, who paid a low fare and embarked in Queenstown. She is misclassified as survivor with mediocre conviction. Her fairness is low since her variables have typical values. Case **c** is also misclassified, but with higher conviction than **b**. This is also a female passenger, but traveling in second class

which made her survival more likely. Point **d** corresponds to a woman traveling in first class. This makes her very likely to survive, hence her high PAC value. Within the casualty class, female first class travelers were rare. Finally, passengers **e** and **f** are a husband and wife traveling third class who paid a low fare, hence they are predicted as casualties. Their high farness is due to the fact that they traveled with 9 parents+children, which is the highest number in the test data.

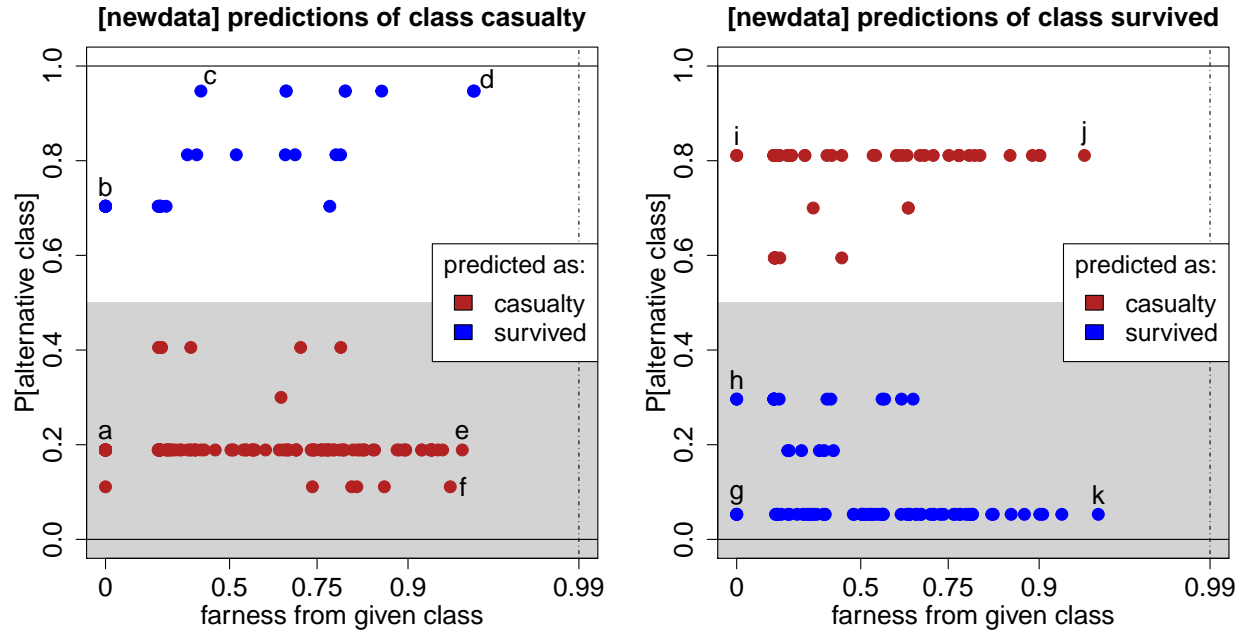


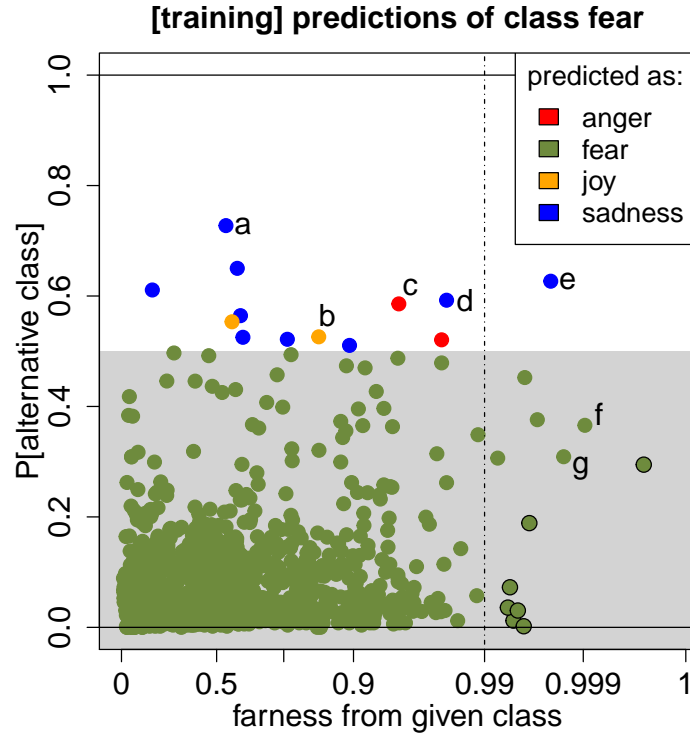
Figure 19: Titanic test data: class maps of classes of casualties (left) and survivors (right).

The class map of the survivors is shown in the right hand panel of Figure 19. Point **g** is a female traveler in second class without any unusual features, so the point has a low PAC and farness. Case **h** is also a female passenger, but traveling in third class. This causes her to be predicted as survivor with less conviction than **g**. Passenger **i** is a male without special characteristics, and therefore predicted as casualty with low farness. Point **j** is also a male passenger, but he paid a very high fare. This makes him stand out from the majority of passengers in the survived class, explaining his high farness. Finally, passenger **k** is a female traveling in first class. This causes her low PAC, that is, she was assigned to the survivor class with high conviction. Her farness is due to paying the highest fare in the test data.

A.5 More on the emotion data

In subsection 4.3 we discussed the classes anger and joy. Here we will address the two remaining classes.

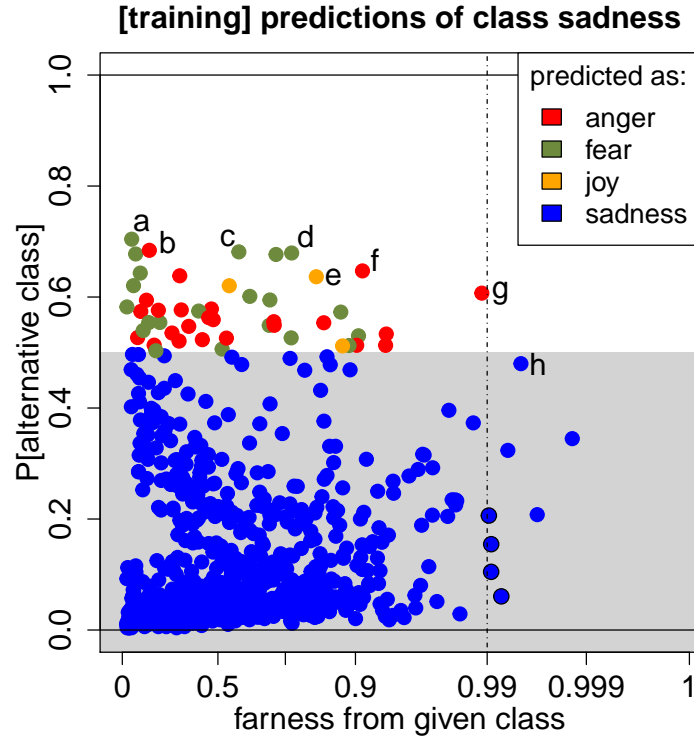
Figure 20 contains the class map of the fear class. Most points are classified correctly as fear, and have unexceptional farness meaning that they sit well within the class. Many of the misclassified points are blue, indicating some confusion with the sadness class. Point **a** is assigned to sadness due to the word ‘lost’. Tweet **b** is predicted as joy due to ‘smile’, but with $PAC(i)$ only slightly above 0.5 (i.e. not with great conviction) due to the word ‘fearing’. Tweet **c** also has a borderline PAC. It is predicted as anger, whereas ‘bully’ is associated with the fear class. Tweets **d** and **e** are predicted as sadness due to the words ‘serious’, ‘sadness’ and ‘despair’, and it is not clear why they were labeled as fear in the first place. The remaining marked points are assigned to fear, their given class. Tweets **f** and **g** contain the words ‘shocking’ and ‘awful’ which are associated with fear. However, they still have an elevated PAC because of the words ‘bitter’ and ‘hilarious’ which are atypical for the fear class. They contain several rare n-grams in the vocabulary such as ‘think they’ or ‘do what’, which increased their farness.



label	tweet
a	I lost my blinders
b	@TheMooseAngel He looks down at his brother, a smile forming on his face. 'What? People fearing me?'
c	@aroseblush Hello !The bigger the bully, the more crocodile tears. Bullies always act like offended victims.
d	@LakersTakeover it ain't that serious. #HOUvsNE #igotbetterthingstodotonight-thandie
e	When the sadness leaves you broken in your bed, I will hold you in the depths of your despair, and it's all in the name of love 🎵
f	#BB18 Michelle crying again #shocking #bitter He's just not that into you 😭 #TeamNicole
g	I really hate Mel and Sue. They think they're hilarious and they're just awful

Figure 20: Class map of the fear class, with the corresponding tweets.

Finally, we discuss the sadness class map presented in Figure 21. The majority of the points are blue, so they were predicted correctly. However, there are quite a few borderline cases with a PAC value somewhat above 0.5. Most of these are predicted as fear or anger, emotions that in some sense lie closer to sadness than joy does. Tweet **a** is short and does not contain enough relevant information. The word ‘despondent’ was too rare to make the vocabulary, so the tweet is predicted in the largest class (fear). Tweet **b** is a borderline case, as the words ‘frown’ and ‘down’ are associated with both anger and sadness. Tweet **c** is predicted as fear, but its label should probably be anger, rather than sadness or fear. It is predicted as fear due to the word ‘shocking’. The words ‘dismal’, ‘useless’, and ‘worst’ point to anger, but they are quite rare in the data and also appear in the fear class. Tweet **d** contains ‘awful’ and ‘anxiety’, causing it to be classified as fear. The classification is not with very high conviction though, due to the word ‘depression’ pointing to sadness. Tweet **e** is a quote and doesn’t have a clear emotion connected to it. The classifier picks up on the word ‘optimism’ which is strongly associated with joy. Tweets **f** and **g** are predicted as anger due to the words ‘anger’ and ‘bitter’. Tweet **h** is a boundary case, containing words pointing to sadness and others to joy.



label	tweet
a	@Christy_RTR @doge_e_fresh I'm despondent
b	@cburt43 turn that frown upside down
c	@Fly_Norwegian quite simply the #worst #airline #worstairline I've ever used! #shocking #appauling #dire #dismal #beyondajoke #useless
d	bad news fam, life is still hard and awful #depression #anxiety #atleastIhaveBuffy
e	"Optimism may sometimes be delusional, but pessimism is always delusional." – Alan Cohen #believe
f	Some moving clips on youtube tonight of the vigil held at Tulsa Metropolitan Baptist church for #TerenceCruther #justice #anger #sadness
g	@FatedDancer '~together.' Hermione lowered her voice slightly, sounding somewhat bitter, perhaps even rueful. 'That would only get you~
h	@BlurtAlerts 'the darkest of nights can be bright, the solemn of faces lights up with a smile'. -@Totemprince believe in me, as I do in you

Figure 21: Class map of the sadness class, with the corresponding tweets.

References

- Raymaekers, J. and P. J. Rousseeuw (2020). *Package cellWise: Analyzing Data with Cell-wise Outliers*. CRAN, R package. <https://CRAN.R-project.org/package=cellWise>.
- Raymaekers, J. and P. J. Rousseeuw (2021b). Transforming variables to central normality. *Machine Learning*, <https://doi.org/10.1007/s10994-021-05960-5> (open access).