Assignment 2

Tutors: **Zhuozhuo Tu, Liu Liu**

Group members: **Chen Chen (cche5002) 480458339,**
Yutong Cao (ycao5602) 470347494,
Yixiong Fang ***

Abstract

The abstract paragraph should be indented 1/2 inch (3 picas) on both left and right-hand margins. Use 10 point type, with a vertical spacing of 11 points. The word **Abstract** must be centred, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

This report documents our modification of support vector machine (SVM) to improve its performance again label noise.

Label noises are unavoidable in many data sets. In probability theory and machine learning, the law of large numbers states that the average of samples drawn converge to the expected value of a random variable with a larger number of trials [Härdle and Simar, 2007]. Besides the choices of hypothesis class, objective function, optimisation method, and output hypothesis, the size and quality of the data used are also crucial in training a model.

Most of the data labelling work are done by humans. Thus mistakes can hardly be avoided, especially when the data size and the number of classes are relatively large. Usually there is a trade-off between the data complexity and the data quality. Weakly supervised learning methods are introduced to handle this kind of problems. Positive and unlabelled (PU) learning and semi-supervised learning are designed to trade data complexity for data quality. They both make use of a small set of correct data to train a model. Also there is learning with noisy labels, on the contrary, trades data quality for data complexity. It learns a model with a large amount of noisy data. In this assignment, we only focused on the latter.

We implemented three methods for learning with noisy labels and conducted experiments with them on two noisy data sets. For both data sets, the flip rates are given. Our first method is to learn with a modified support vector machine. The label noise information is added to the optimisation term of the SVM as a function of the observed label and an error term which follows a Bernoulli distribution. This method is first proposed by Biggio et al. [2011] to learn with random classification noise (RCN), and we extended it to learn with class-dependent noise (CCN). The second method is the heuristic approach based on the work of Wu et al. [2003]. This method follows the assumption that models trained with the noisy data still give an adequate performance. Thus the predicted probability of a noisy sample point belonging to one class should be around 0.5 as the model can neither accept nor reject the prediction with confidence. We exclude the potential noisy data points following this idea and train a model on the rest data which we believe are cleaner than the original data. The third method is the importance reweighting approach proposed by Liu and Tao [2016]. This method calculates a reweighting coefficient for each sample point using the result from training a model on the original noisy data to reweight the loss for each sample point. For all three methods, we choose use the same classification model so that we can compare the three learning with label noise methods with the rest conditions controlled. Although SVM is not considered to be robust to label noise, its classification accuracy is relatively high compared with other models.

We used the three methods to learn two sets of noisy data, both containing 10,000 instances belongs to two classes. The first data set is sampled from the fashion-mnist database and the second is sampled from the CIFER database. We will compare our results from using the three learning with label noise methods on both data sets in accuracy and process time. For all three methods, the values of noise rates are required. We implemented a noise rate estimation using the method proposed in the work of Liu and Tao [2016]. We did not use the estimated values in our classification experiments as the true rates are provided. But we will compare our estimation with the true values.

2 Related work

In the literature, there are many works on learning with label noise. One popular approach is to use classification algorithms that are proved to be robust to label noise. Frénay and Verleysen [2014] summarised in their work that 0-1 loss and least-squares loss are robust loss functions to label noise. Also, bagging and ensemble decision tree are robust classification models to learn with label noise. This kind of approach does not deal with label noise directly. Instead, they treat the noisy sample points as outliers and assume they will not have too much effect on the trained model. We chose not to use the robust loss functions as the corresponding classification models may not produce satisfying classification accuracy as SVM. But we used a bagging-like technique by sampling multiple times and averaging the predicted probabilities to generate better and more stable predictions. Frénay and Verleysen [2014] also mentioned filtering methods for learning with label noise. This kind of methods try to remove mislabelled data before training a final model. Our second model follows this idea by excluding data points with vague predictions. Yang et al. [2018] also proposed a filter-like method that filters the data during training and calculates a mislabelled probability for each sample rather than removing them directly. Many of the approaches mentioned above do not consider the noise rates as known. Although this makes the problem general, the methods may not be able to handle more complicated problems.

Biggio et al. [2011] proposed a method that rewrites the objective function and adds the noise rate information to it. They substitutes the given label values with expected values and optimises the new objective function. This is what our first learning with label noise method is based one. The idea behind this method is clear and intuitive, but it is only applicable for certain loss functions. Liu and Tao [2016] proposed an importance reweighting method that reweights the surrogate loss function using information from a pre-train process. This method does not have limitations on the type of surrogate functions chosen. And they also provided an efficient method for estimating the noise rate.

3 Method

We use SVM with Gaussian kernel as the base classification method for this task because of its generalisation ability.

SVM with Gaussian kernel was first published by Boser et al. [1992]. Cortes and Vapnik [1995] an improvement with soft-margin SVM to avoid over-fitting problem. The SVM is to minimise the Hinge loss function

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(w \cdot x_i - b)\right)\right] + \lambda\|w\|^2.$$

Here, w is the weighting factor, b is a constant. We classify the ith image into category 1 or -1 when $w \cdot x_i - b \geq 1$ or $w \cdot x_i - b \leq -1$, respectively. This loss function not only penalises points that misclassified, but also points that are closed to the dividing hyperplane, with a regularisation term $\lambda\|w\|^2$. Fernández-Delgado et al. [2014] suggested SVM is very likely to be one of the most powerful classifiers, by comparing 179 classification algorithms over 121 large data sets. The distances between data points and the dividing hyperplane gives an intuitive estimate of the generalisation ability of the trained SVM model [Hastie et al., 2001].

In this image classification assignment, we do not observe the true labels y_i. This section proposes three different approaches to modify ordinary SVM to attack label noise problem. However, we do not have access to a test data set with true labels (x_i, y_i) to verify the generalisation ability of our three different models, which are all trained by noisy data (x_i, S_i). Using SVM, the gap between hyperplane and the training data set provides a natural and 'free' metric of its generalisation ability [Hastie et al., 2001], and also provides an alternative to using a test data set. As of the strong generalisation ability of a well-trained SVM [Cortes and Vapnik, 1995, Jin and Wang, 2012], it was chosen as our classification method.

3.1 Preprocess

`StandardScaler` from `sklearn` was used. Each image is rescaled to have mean zero and standard deviation one. By doing this, we removed the brightness difference among different images. This process is called photometric normalisation. Jonsson et al. [2002] suggested that this may improve the performance of Gaussian kernel SVM as Gaussian kernel is a radius based kernel and hence it performs well when images are on the same scale. For the CIFER dataset, we used principle component analysis (PCA) for dimension reduction. We did not run PCA on the fashion-mnist data because the dimension is reasonable.

3.2 The original data set is balanced

The assignment instruction states the probabilities $P(S = 1|Y = 0) = 0.2$ and $P(S = 0|Y = 1) = 0.4$. From the data, we observed that 40% of contaminated labels S (i.e. $P(S = 1)$) is 1. As a result

$$\begin{aligned} P(S=1) &= P(S=1|Y=1)P(Y=1) + P(S=1|Y=0)P(Y=0) \\ &= [1 - P(S=0|Y=1)]\, P(Y=1) + P(S=1|Y=0)P(Y=0) \\ &= 0.6P(Y=1) + 0.2\,[1 - P(Y=1)] = 0.4, \end{aligned}$$

which implies $P(Y = 1) = P(Y = 0) = 0.5$ and the original classification problem is balanced.In addition, define the Bernoulli random variable $\epsilon(S)$ with the means $E(\mu(\epsilon)(S = 0)) = P(Y = 1|S = 0) = 0.5 \times 0.4/0.6 = 1/3$ and $E(\epsilon(S = 1)) = P(Y = 0|S = 1) = 0.5 * 0.2/0.4 = 0.25$. This random variable ϵ describe the unobserved random label noise. Hence the expectation is

$$E\epsilon(S) = P(Y=0|S=1)P(S=1) + P(Y=1|S=0)P(S=0) = 0.25 \times 0.4 + 1/3 \times 0.6 = 0.3. \quad (1)$$

3.3 Method 1: Modified support vector machine

We proposed a modified support vector machine (SVM) to attack this classification problem with label noise. We describe the mathematical justification for our modification.

3.3.1 Expectation maximisation

This section extends the expectation maximisation algorithm proposed by Biggio et al. [2011]. The original algorithm was proposed to encounter random classification noise where the flip rate $\rho_+ = \rho_-$ and we extend it to manage the case dependent label noise where the flip rates $\rho_+ \neq \rho_-$.

Recall the dual problem of an SVM is to maximise

$$f(c_1 \dots c_n) = \sum_{i=1}^{n} c_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i c_i k(x_i, x_j) y_j c_j, \quad (2)$$

subject to $\sum_{i=1}^{n} c_i y_i = 0$, and $0 \leq c_i \leq \frac{1}{2n\lambda}$ for all i. Here y_i and x_i are the labels and features of the ith image, c_i is the ith Lagrangian multiplier, $k(x_i, x_j)$ is the Gaussian kernel product of x_i and x_j

$$\exp(-\gamma \| x_i - x_j \|).$$

Substitute label noise $y_i = S_i(1 - 2\epsilon(S_i))$ into objective function (2)

$$f(c_1 \dots c_n) = \sum_{i=1}^{n} c_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} S_i c_i k(x_i, x_j) S_j c_j (1 - 2\epsilon(S_i))(1 - 2\epsilon(S_j)), \quad (3)$$

When $i = j$, the expectation $E(1 - 2\epsilon(S_i))(1 - 2\epsilon(S_j)) = 1 - 4E\epsilon(S_j) + 4E\epsilon(S_j^2) = 1$. When $i \neq j$, the expectation $E(1 - 2\epsilon(S_i))(1 - 2\epsilon(S_j)) = (1 - 2E\epsilon(S_i))(1 - 2E\epsilon(S_j)) = 1 - \mu$ where parameter $\mu := 0.84$, by substituting expectations (1) from Section 3.2.

Define kernel correction matrix M with the (i,j)-th entry being m_{ij}.The diagonal entries $m_{ii} = 1$, and the off diagonal entries $m_{ij} = 1 - \mu = 0.16$ when indexes $i \neq j$. Using the technique of Expectation maximisation,

$$Ef(c_1 \dots c_n) = \sum_{i=1}^{n} c_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} S_i c_i k(x_i, x_j) S_j c_j m_{ij}, \tag{4}$$

The only difference between our modified SVM and ordinary SVM is to replace the kernel matrix K with our new proposed matrix $Q := K \circ M$.

3.3.2 Tuning

Define data vector $\vec{x} := (x_1, x_2, \dots, x_n)$. The Kernel parameter for Gaussian Kernel γ is chosen to maximise the Variance of Kernel matrix $K(\vec{x}, \vec{x})$. The regularisation parameter is chosen to be one as the model seems to be insensitive to the regularisation parameter.

This method gives us an accuracy of 94.6% on the testing data.

3.4 Method 2: heuristic approach

Method 2 still implement SVM. We chose SVM because classification algorithms with Hinge loss, including SVM, is robust against random classification label noise. Our label noise is class dependent. However, experiment shows that SVM is still robust against it.

3.4.1 Select samples

Ordinary SVM only gives a classification without revealing a probability that indicates the confidence of classification. Wu et al. [2003] proposed a five-fold cross-validation method to calculate the classification probabilities $P(Y = 1|X_i)$ for SVM. Using this method, we calculated the probability $P(Y = 1|X_i)$ with label noise by an SVM with Gaussian kernel. We have $10,000$ samples. We only use those with largest and smallest $P(Y = 1|X_i)$ (first $1/3$ and last $1/3$), because intuitively the contaminated samples are more likely to have a probability $P(Y = 1|X_i)$ close to 0.5 and the error rate $P(\epsilon = 1) = 0.3$. A 3% margin was left because classification errors must exist.

3.4.2 Label correction

We relabel the $1/3$ of the samples with highest fitted probability $P(Y = 1|X_i)$ as 1 and the $1/3$ samples with the smallest fitted probability $P(Y = 1|X_i)$ as -1. This step is important because section 3.2 shows that the original data set is balanced ($P(Y = 1) = P(Y = 0) = 0.5$).

Training our SVM with Gaussian Kernel again with this subset of relabeled data gives our second model.

This method gives us an accuracy of 95.0% on the testing data.

3.5 Method 3: Reweighting

Please copy some formulae to here from tut11.

4 Result

scatter plot of time vs. accuracy histogram of accuracy mean, sd table

Please pay special attention to the instructions in section 6 regarding figures, tables, acknowledgments, and references.

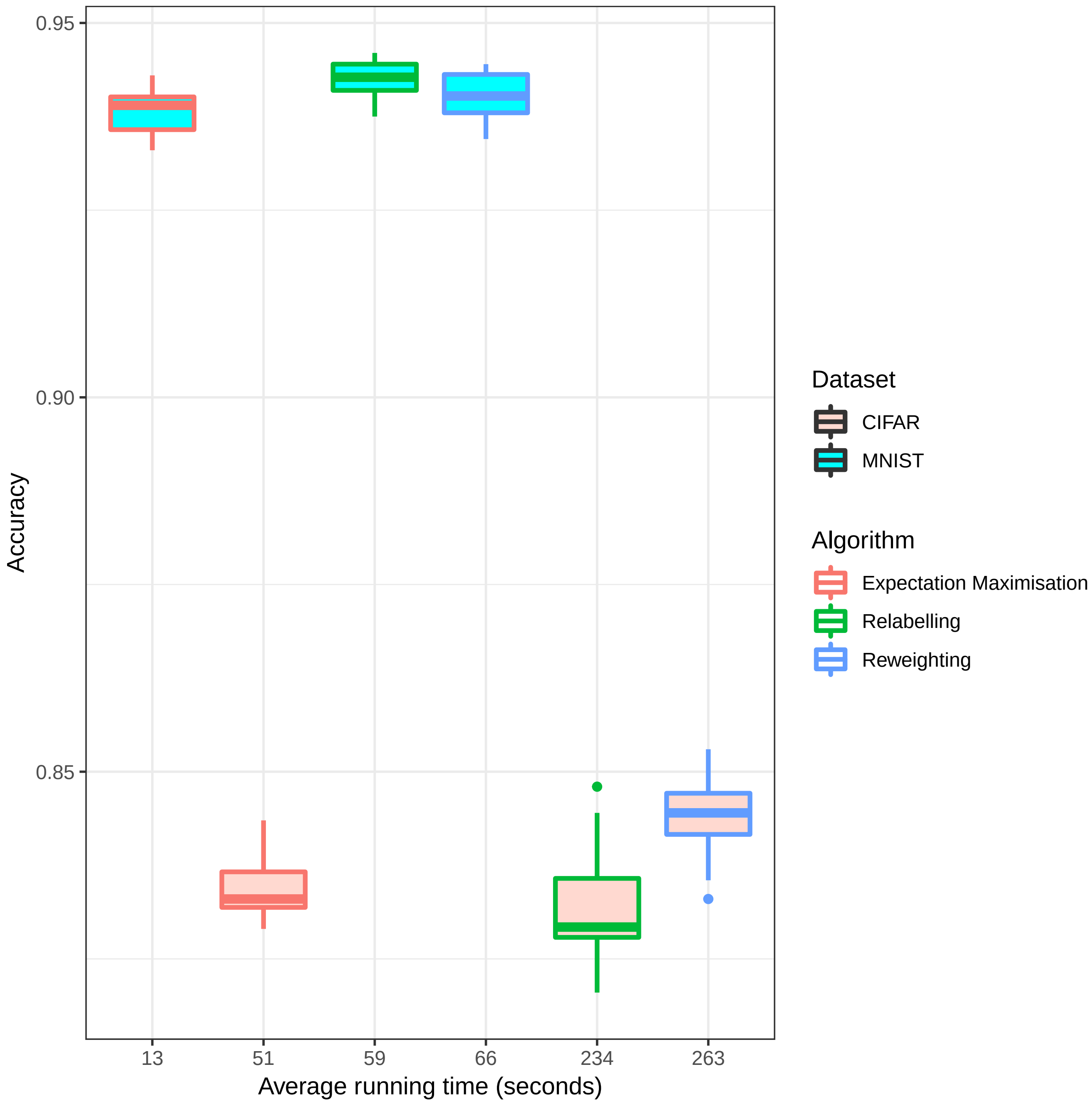

Figure 1: Boxplot

Data	Null Hypothesis (H_0)	D	P-value	Reject H_0
MNIST	Relabelling algorithm is no more accurate than expectation maximisation.	0.625	0.0019	Reject
	Expectation maximisation algorithm is as accurate as Reweighting.	0.3125	0.4154	Fail to reject
	Reweighting algorithm is as accurate as relabelling.	0.3125	0.4154	Fail to reject
CIFAR	Expectation maximisation algorithm is no more accurate than relabelling.	0.5	0.0183	Reject
	Reweighting algorithm is no more accurate than expectation maximisation.	0.6875	0.0005	Reject
	Reweighting algorithm is no more accurate than relabelling.	0.6875	0.0005	Reject

Table 1: Hypothesis test

Table 2: Mean sd

Mean	Standard deviation	Confidence interval
0.835	0.004	(0.833.0.837)

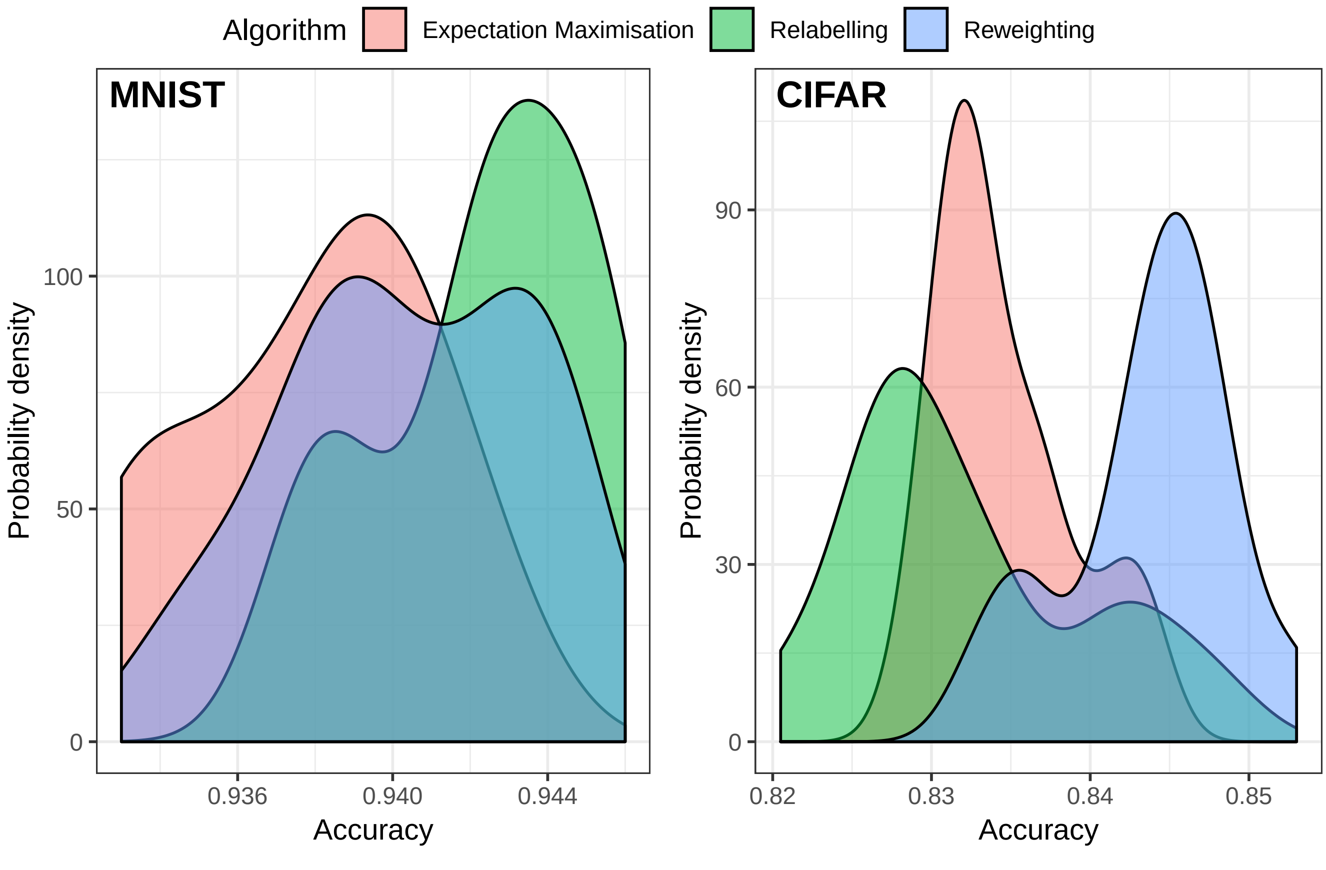

Figure 2: Density function from kernel smoothing

5 Headings: first level

First level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 12. One line space before the first level heading and 1/2 line space after the first level heading.

5.1 Headings: second level

Second level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 10. One line space before the second level heading and 1/2 line space after the second level heading.

5.1.1 Headings: third level

Third level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 10. One line space before the third level heading and 1/2 line space after the third level heading.

6 Citations, figures, tables, references

6.1 Citations within the text

Citations within the text should be numbered consecutively. The corresponding number is to appear enclosed in square brackets, such as [1] or [2]-[5]. The corresponding references are to be listed in the same order at the end of the paper, in the **References** section. (Note: the standard BibTeX style `unsrt` produces this.) As to the format of the references themselves, any style is acceptable as long as it is used consistently.

6.2 Footnotes

Indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).[2]

6.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction; art work should not be hand-drawn. The figure number and caption always appear after the figure. Place one line space before the figure caption, and one line space after the figure. The figure caption is lower case (except for first word and proper nouns); figures are numbered consecutively.

Make sure the figure caption does not get separated from the figure. Leave sufficient space to avoid splitting the figure and figure caption.

You may use color figures. However, it is best for the figure captions and the paper body to make sense if the paper is printed either in black/white or in color.

6.4 Tables

All tables must be centered, neat, clean and legible. Do not use hand-drawn tables. The table number and title always appear before the table. See Table 3.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

[1] Sample of the first footnote

[2] Sample of the second footnote

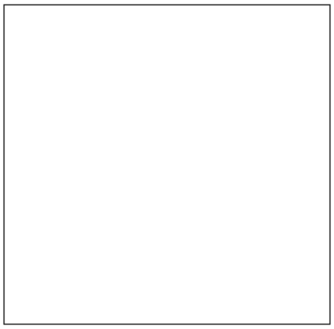

Figure 3: Sample figure caption.

Table 3: Sample table title

PART	DESCRIPTION
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

6.5 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the graphicx package. Always specify the figure width as a multiple of the line width as in the example below using .eps graphics

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.eps}
```

or

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

for .pdf graphics. See section 4.4 in the graphics bundle documentation (`http://www.ctan.org/tex-archive/macros/latex/required/graphics/grfguide.ps`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command.

References

Wolfgang Härdle and Léopold Simar. *Applied multivariate statistical analysis*, volume 22007. Springer, 2007.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Support vector machines under adversarial label noise. In *Proceedings of the Asian Conference on Machine Learning*, volume 20, pages 97–112. PMLR, 2011.

Tingfan Wu, ChihJen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2003.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2016.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2014.

Pengyi Yang, John T Ormerod, Wei Liu, Chendong Ma, Albert Y Zomaya, and Jean YH Yang. Adasampling for positive-unlabeled and label noise learning with bioinformatics applications. *IEEE Transactions on Cybernetics*, (99):1–12, 2018.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM. ISBN 0-89791-497-X. doi: 10.1145/130385.130401. URL http://doi.acm.org/10.1145/130385.130401.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995. ISSN 1573-0565. doi: 10.1007/BF00994018.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.*, 15(1):3133–3181, January 2014. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2627435.2697065.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Chi Jin and Liwei Wang. Dimensionality dependent pac-bayes margin bound. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1034–1042. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4500-dimensionality-dependent-pac-bayes-margin-bound.pdf.

Kenneth Jonsson, Josef Kittler, YP Li, and Jiri Matas. Support vector machines for face authentication. *Image and Vision Computing*, 20(5-6):369–375, 2002.