UDC 004.855.5(045)

V.V. Kalmykov, V.M. Sineglazov
(National Aviation University, Ukraine)

Overview on semi-supervised learning methods

Proposed a new approach to semi-supervised learning in aviation. Provided the results of semi-supervised methods modelling.

Semi-supervised learning.

Semi-supervised learning is a wide category of machine learning techniques that use both labelled and unlabelled data; thus, as the name implies, it is a hybrid technique between supervised and unsupervised learning.

In general, the basic idea behind semi-supervision is to treat a data point differently depending on whether it is labelled or not: for labelled points, the algorithm uses traditional control to update model weights; and for unlabelled points, the algorithm minimizes the difference in predictions between other similar training samples.

For example, we will consider a binary classification problem Figure 1. Suppose we have only 12 labelled data points, and the rest are unlabelled.

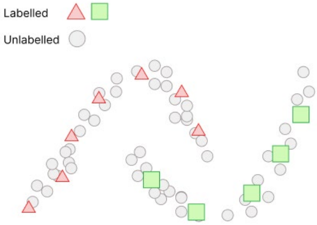


Fig.1. Dataset example for a binary classification problem

Supervised learning updates the model weights to minimize the average difference between predictions and labels. However, with a limited amount of labelled data, this can lead to a decision boundary that is fair for the labelled points, but does not generalize to the entire distribution-as in Figure 2 below.

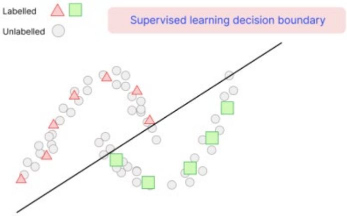


Fig.2. Supervisedlearningdecisionboundary

On the other hand, unsupported learning attempts to cluster points based on similarity in some feature space. But without labels to guide the training, the unsupervised algorithm may find suboptimal clusters. In Figure 3, for example, the clusters found do not correspond correctly to the true class distribution.

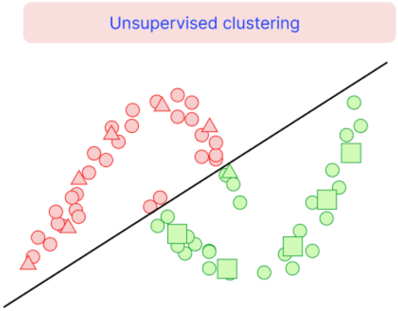


Fig.3. Unsupervisedclustering

Without enough labelled data or in complex clustering conditions, supervised and unsupervised methods may not produce the desired results.In the semi-supervised method, however, we use both labelled and unlabelled data. Our labelled data points act as a sanity check; they justify the predictions of our model and add structure to the learning problem by determining how many classes exist and which clusters correspond to which class.Unlabelled data points provide context; by exposing our model to as much data as possible, we can accurately estimate the shape of the entire distribution.By having both parts, labelled and unlabelled data, we can train more accurate and robust models. In our dataset, training with semi-supervised approach allows us to approximate the true distribution shown in Figure 4.

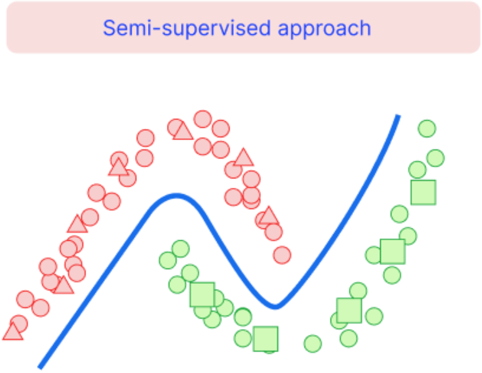
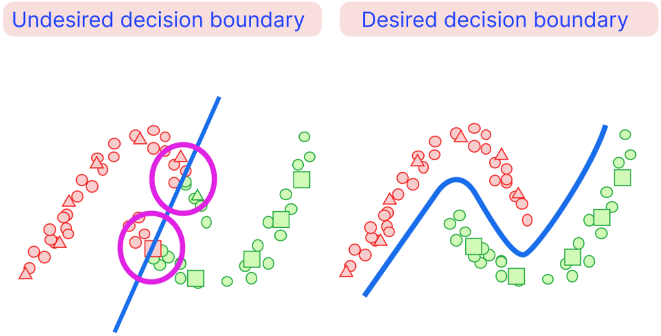


Fig. 4. Semi-supervised approach

The Cluster, Consistency, and Manifold Assumptions

As a broad subset of machine learning, semi-supervised intuition is based on several basic principles. The continuity, or smoothness, assumption indicates that closely spaced data points are more likely have the same label.

Similarly, the clustering assumption indicates that in a classification problem, the data tends to organize into clusters of high density, and that data points in the same cluster are likely to have the same label. Therefore, the decision boundary should not lie in areas with dense packing of data points; rather, it should lie between areas of high density, dividing them into discrete clusters.



Undesired decision boundary

Desired decision boundary

Fig. 5. Undesired and desired decision boundary

The manifold assumption adjustspredictions for our data to deep learning applications, including natural language processing and computer vision. It suggests that the distribution of high-dimensional data can be represented in an embedded low-dimensional space.This low-dimensional space is called the data manifold.

Consider the problem of binary classification of images of cats and dogs. In deep learning applications, an image is simply a large tensor of values indicating pixel colors. This image space is our multidimensional space.

Based on values of colors, images of dogs and horses are dispersed in an obscure distribution in anEuclidean high-dimensional space, in which, there is no clear clusters. Therefore, it is possible to assume that exists a lower-dimensional manifold such that the idea of distance is representative of semantic meaning—in this case, where datapoints of dogs are clustered near dogs, and horses are clustered near horses.

In deep learning, working with the distribution of high-dimensional images of dogs and horses is tough. That is why, based on the manifold assumption, our model can learn the function mapping images in Euclidean space to representations on our low-dimensional manifold. So that, our cluster and continuity assumptions are more reliable, and we can classify a datapoint based on its learned representation.

The manifold assumption helps to harness semi-supervised techniques in deep learning settings.

Semi-supervised learning techniques

The main motivation for using consistency regularization is to take advantage of continuity and cluster assumptions.

In a semi-supervised environment, suppose we have a dataset with labeled and unlabeled examples of two classes.

During training, we treat labeled and unlabeled data points differently: for labeled points, we optimize using traditional supervised learning, calculating losses by comparing our prediction to our label; for unlabeled points, we want to ensure that on our low-dimensional manifold, similar data points have similar predictions. To enforce consistency let us consider dataset D so that:

$$D = \{(X_{labeled}, Y), (X_{unlabled})\}, Y \ = \ \{y_1, y_2, \dots\}, y_i \in \{green, red\} \qquad (1)$$

With different augmentation techniques, we can artificially create similar datapoints.

Consider the Augment(x) function, which slightly modifies x. We need to make sure that our model produces the same predictions for the augmented data point, Augment(x), and its original counterpart, x.

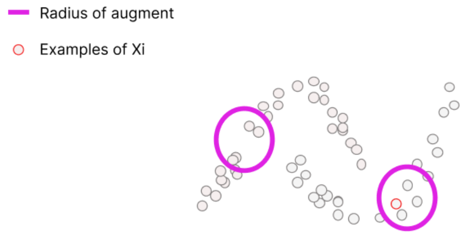


Fig. 6. Augmentation function

For a given image x, our model must make similar predictions for all data points in the radius of the Augment(x). In practice, this is achieved by introducing both a controllable and an uncontrollable loss term. Some of the most popular implementations of consistency regularization are Pi-Models and Temporal Ensembling, proposed by Lane and Ayla [1].

WithCrossEntropy, a supervised function of loss, and the model f Laine and Aila formulate the loss as follows:

$$loss = \begin{Bmatrix} CrossEntropy(x_i, y_i) + ||f_0(x_i) - f_0(x_i) - f_0\big(Augment\ (x_i)\big)||_2^2, x_i \in X_{labled} \\ ||f_0(x_i) - f_0\big(Augment\ (x_i)\big)||_2^2, x_i \in X_{unlabled} \end{Bmatrix} (2)$$

Optimizing this loss for unlabelled data points results in the distance - measured by the L2-norm - between predictions for any Augment(x) must be the same as the prediction for the original x. By minimising the distance between predictions of similar data points, we will find a decision boundary consistent with our continuity and clustering assumptions.

The term unobserved loss directly induces the model to assign similar data points to the same cluster; and if the model predictions agree within a certain radius around each data point x, then the decision boundary will be away from clusters with high data density.

Pseudo-labeling

Pseudo-labelling is a concept where, in process of training, model’s predictions are converted into a “one-hot” label.

$$Model\,prediction\begin{pmatrix}0.65\\0.15\end{pmatrix} \rightarrow Conversion\ to\ the\ "one-hot"\ \ label\begin{pmatrix}1\ green\\0\ blue\end{pmatrix}$$

All the confident predictions of the model are converted into "one-hot" vectors, where the most confident classes become the labels. Based on this, we train on the new one-hot probability distribution as a pseudo-label.

Not only can we create artificial labels, but training on pseudo labels is a form of entropy minimisation, which means that model predictions are encouraged to have high confidence on unlabelled data points. Similarly, by taking certain predictions as true, we avoid learning any general rules about the true distribution of the data (inductive learning). Thus, pseudo-tags offer a form of transductive learning - reasoning from given training data to other specific test data.

References

1. SamuliL., Timo A., Temporal Ensembling for Semi-supervised Learning. URL - https://arxiv.org/pdf/1610.02242.pdf
2. PhilipBachman, OuaisAlsharif, andDoinaPrecup. Learningwithpseudo-ensembles. InAdvancesinNeuralInformationProcessingSystems 27 (NIPS). 2014.
3. KaimingHe, XiangyuZhang, ShaoqingRen, andJianSun. Delvingdeepintorectifiers: Surpassinghuman-levelperformanceonimagenetclassification. CoRR, abs/1502.01852, 2015.
4. DmytroMishkinandJiriMatas. Allyouneedis a goodinit. InProc. InternationalConferenceonLearningRepresentations (ICLR), 2016.
5. Patrice Y. Simard, Yann A. LeCun, John S. Denker, andBernardVictorri. TransformationInvarianceinPatternRecognition — TangentDistanceandTangentPropagation, pp. 239–274. 1998.
6. TheanoDevelopmentTeam. Theano: A Pythonframeworkforfastcomputationofmathematicalexpressions. CoRR, abs/1605.02688, May 2016.
7. XiaojinZhu. Semi-supervisedlearningliteraturesurvey. TechnicalReport 1530, ComputerSciences, UniversityofWisconsin-Madison, 2005.
8. XiaojinZhuandZoubinGhahramani. Learningfromlabeledandunlabeleddatawithlabelpropagation. TechnicalReport CMU-CALD-02-107, CarnegieMellonUniversity, 2002.