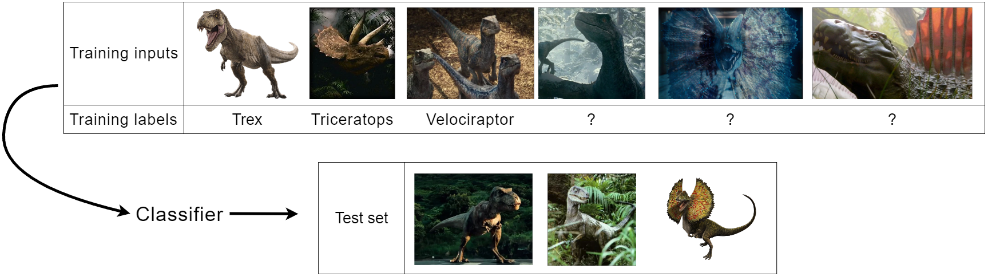
# Semi-Supervised Learning

INFS4203/7203 Tutorial Week 10, Semester 2, 2024

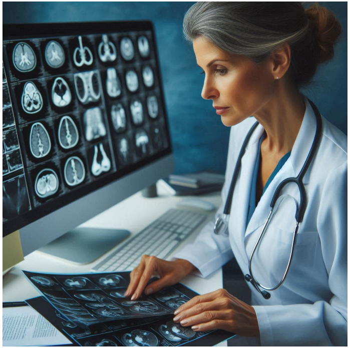
Jonathan Wilton



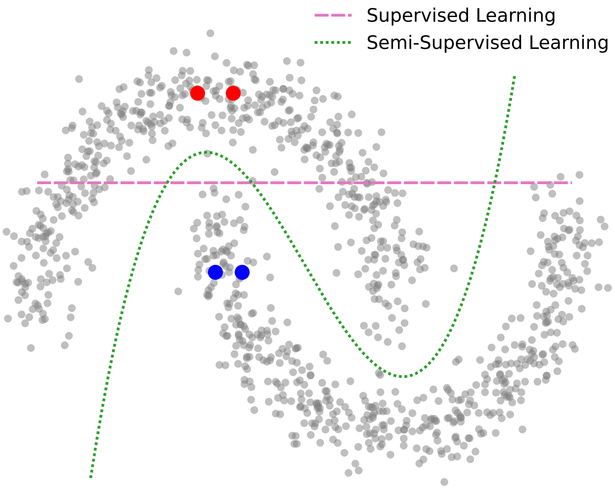
# Motivation

Use difficult to obtain labelled data and easy unlabelled data to train an accurate model.

Examples: medical diagnosis, image captioning (figures generated using Copilot).

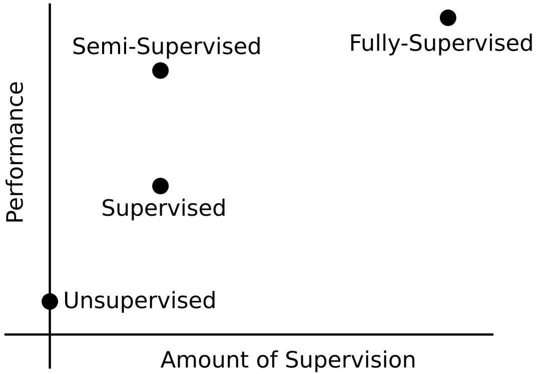


# Semi-supervised learning



# Semi-supervised learning

There is a trade-off between amount of supervision and predictive performance.

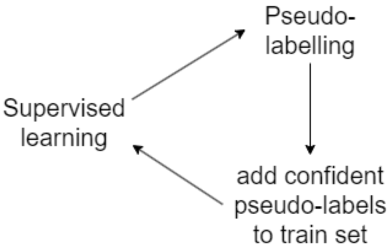


# Your turn

- Can you think of an effective way to use unlabelled data to help train a classifier?

# Example: Pseudo-Labelling

1. Train a classifier using supervised learning
2. While not converged:
   1. Use classifier to predict the labels for unlabelled examples
   2. Add to training set the unlabelled examples with confident predictions
   3. Re-train classifier on the new training set



# Example: Simple Pseudo-Labelling in Python

```python
import numpy as np
from sklearn.datasets import make_moons
from sklearn.semi_supervised import SelfTrainingClassifier
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
np.random.seed(42)

X_tr, X_ts, y_tr, y_ts = train_test_split(*make_moons(1000, noise=0.1), train_size=0.7)

number_of_labels = 4
lb_idx = np.random.choice(np.arange(len(X_tr)), number_of_labels)

clf_supervised = RandomForestClassifier().fit(X_tr[lb_idx], y_tr[lb_idx])
print("Test accuacy supervised:", clf_supervised.score(X_ts,y_ts))

y_ssl = -np.ones_like(y_tr)
y_ssl[lb_idx] = y_tr[lb_idx]
clf_semisupervised = SelfTrainingClassifier(RandomForestClassifier()).fit(X_tr, y_ssl)
print("Test accuracy semi-supervised:", clf_semisupervised.score(X_ts,y_ts))
```

```
Test accuacy supervised: 0.36
Test accuracy semi-supervised: 0.8366666666666667
```

# Example: Pseudo-Labelling

Question: what are some pros and cons of pseudo-labelling?

# Example: Pseudo-Labelling

Question: what are some pros and cons of pseudo-labelling?

Pros:

- simple to implement,
- potential for significantly improved generalisation by utilising unlabelled data,
- compatible with most existing supervised learning algorithms.

# Example: Pseudo-Labelling

Question: what are some pros and cons of pseudo-labelling?

Pros:

- simple to implement,
- potential for significantly improved generalisation by utilising unlabelled data,
- compatible with most existing supervised learning algorithms.

Cons:

- pseudo-labels can be incorrect and lead to confirmation bias,
- quantity-quality tradeoff can be difficult to tune.

# Example: Consistency Regularisation

Assume that similar inputs should be assigned similar outputs.

Graph regularisation:


Data augmentation:

# Example: Consistency Regularisation

Pros:

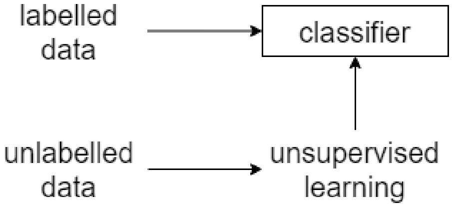
- Label information can be propagated to unlabelled data $\rightarrow$ improved generalisation,
- Assumption often holds in practice.

Cons:

- Can be more difficult to implement than pseudo-labelling,
- Assumption may not hold $\rightarrow$ worse generalisation (e.g., adversarial examples, poor quality augmentations).

# Example: Unsupervised Pre-Training

Use labelled data + unsupervised learning on the unlabelled data to train the model.




Examples: cluster-then-label, representation learning, unsupervised warm-up.

# Example: FixMatch + Extensions (Advanced)

Combine pseudo-labelling with consistency regularisation [Sohn et al., 2020]:

$$\frac{1}{n_L}\sum_{i=1}^{n_L} \mathrm{Loss}(y_i, f(\omega(\boldsymbol{x}_i))) + \frac{1}{n_U}\sum_{i=1}^{n_U} \mathbb{1}(\max(q_i) \geq \tau)\,\mathrm{Loss}(\hat{q}_i, f(\Omega(\boldsymbol{u}_i))).$$

Threshold $\tau$ controls the quantity-quality trade-off.

# Example: FixMatch + Extensions (Advanced)

Combine pseudo-labelling with consistency regularisation [Sohn et al., 2020]:

$$\frac{1}{n_L}\sum_{i=1}^{n_L}\mathrm{Loss}(y_i, f(\omega(\boldsymbol{x}_i))) + \frac{1}{n_U}\sum_{i=1}^{n_U}\mathbb{1}(\max(q_i) \geq \tau)\,\mathrm{Loss}(\hat{q}_i, f(\Omega(\boldsymbol{u}_i))).$$

Threshold $\tau$ controls the quantity-quality trade-off.

Some Extensions:

- Fairness/uniform alignment (FreeMatch) [Wang et al., 2023],
- Adaptive and data-dependent threshold (SoftMatch) [Chen et al., 2023],
- Learn to correct the noisy pseudo-labels (InstanT) [Li et al., 2023],

# Future Research Directions

- Theoretical guarantees for semi-supervised learning methods.
- Can we get good performance with weaker assumptions?
- Closing the performance gap to fully-supervised learning.
- Most current research focuses on neural-networks – extend to other models.

# Bigger Picture

Some other related weakly-supervised learning problems:

- Semi-supervised regression,
- few shot learning,
- positive-unlabeled learning,
- learning with noisy labels,
- complementary label learning.

# Check your understanding

- What is semi-supervised learning?
- Applications of semi-supervised learning.
- Example methods, assumptions, limitations.
- Other weakly-supervised learning problems.

Further reading:

- sklearn documentation,
- survey paper [Engelen and Hoos, 2020],
- textbook [Chapelle et al, 2006],
- research papers.

Link to resources:

https://github.com/jonathanwilton/DMWK10