Default Project Guidelines COMP6321

The goal of this project is to give students experience in the following aspects of applied machine learning:

- collecting data sets that are given in diverse formats;
- preparing data sets for training and testing;
- training many types of classification and regression models;
- performing hyperparameter search for each type of model;
- evaluating and summarizing classification and regression performance;
- generating plots that summarize performance;
- different approaches to interpretability; and
- writing about machine learning concepts, experiments, and results.

The specifics of these skills are not spelled out in the project guidelines, but rather something you should acquire by drawing upon lectures, labs, papers, lots of practice, and feedback from course staff.

Project Overview

The title and abstract of the report template (`report.tex` or `report.pdf`) provides high-level guidance of what your project should aim to achieve. Read that document now, before continuing with the guidelines below.

Be prepared to invest a little time in understanding and processing the datasets. This can be tedious, and not all of the datasets are described very well, but keep in mind that these datasets have already been prepared for machine learning. Cleaning and processing datasets in the real world (in industry, in science) is often much harder and more time consuming than the datasets here!

1. Classification

The idea here is to train and evaluate 8 classification methods across 10 classification datasets.

Classification models

You should evaluate the following classification models:

- *k*-nearest neighbours classification
- Support vector classification
- Decision tree classification
- Random forest classification
- AdaBoost classification
- Logistic regression (for classification)
- Gaussian naive Bayes classification

- Neural network classification

Each of these is provided by scikit-learn under a unified interface. For example, MLPClassifier implements a fully-connected neural network classifier (also called a multi-layer perceptron, or MLP), and GaussianNB implements a Gaussian naive Bayes classifier. The AdaBoostClassifier implements AdaBoost for classification, for which using the default *base_estimator* is OK to use. Even though a model like logistic regression is not strictly a classifier, the scikit-learn implementation will still predict class labels.

Hyperparameters. Some types of models have more hyperparameters than others. You do not need to try every hyperparameter. Just choose 1–3 hyperparameters that are likely to have impact, such as C and γ for SVM, or *max_depth* and *n_estimators* for random forests, or *hidden_layer_sizes* and *learning_rate* and *max_iter* for neural networks. You need to choose and justify your strategy for picking hyperparameter ranges and for sampling the hyperparameters during hyperparameter search, and you should specify how you trained your final model once the best hyperparameters were found.

Classification datasets

You should evaluate each of the above classification model families on each the following UCI repository datasets:

1. Diabetic Retinopathy
2. Default of credit card clients
3. Breast Cancer Wisconsin
4. Statlog (Australian credit approval)
5. Statlog (German credit data) (recommend `german.doc` for instructions and `german-numeric` for data.)
6. Steel Plates Faults
7. Adult
8. Yeast
9. Thoracic Surgery Data
10. Seismic-Bumps

For these datasets you'll need to read the data set descriptions and discern which fields are intended to be features and which are the class labels to be predicted. If a dataset does not come with an explicit train/test split, then you will have to ensure your methodology can still evaluate the performance of the model on held-out data. Your conclusions regarding classification should draw from training and evaluating on the above datasets.

2. Regression

The idea here is to train and evaluate 7 regression methods across 10 regression datasets.

Regression models

You should evaluate the following regression models:

- Support vector regression
- Decision tree regression
- Random forest regression
- AdaBoost regression
- Gaussian process regression
- Linear regression
- Neural network regression

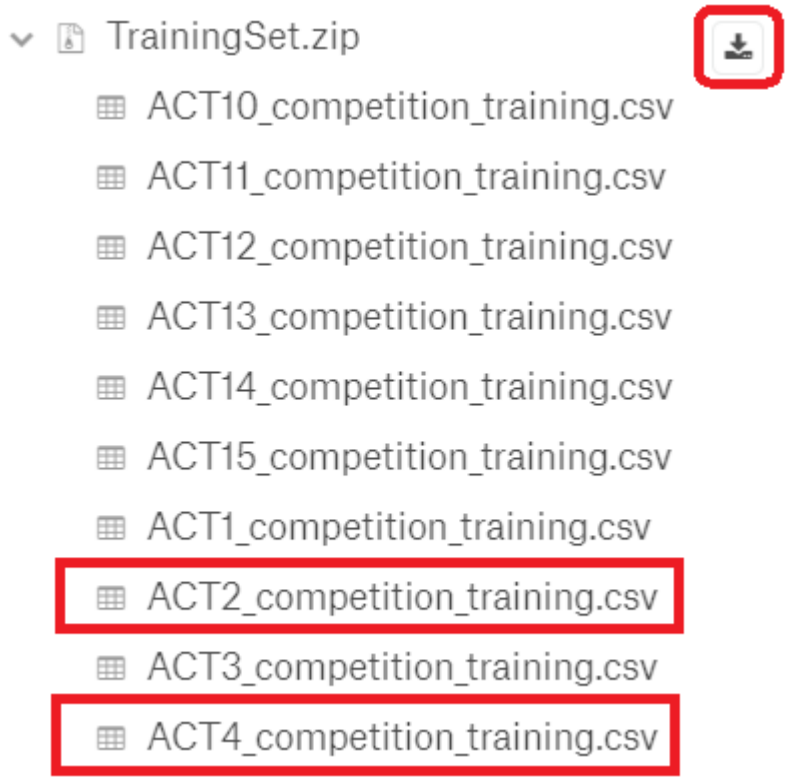
Each of these is provided by scikit-learn. For example, the MLPRegressor class implements a fully-connected neural network for regression. The SVR class implements support vector regression. The AdaBoostRegressor implements AdaBoost for regression, for which using the default *base_estimator* is OK. The GaussianProcessRegressor implements Gaussian process regression.

Regression datasets

You should evaluate each of the above regression model families on each the following datasets, which are mostly again from the UCI repository:

1. Wine Quality
2. Communities and Crime
3. QSAR aquatic toxicity
4. Parkinson Speech (extract RAR files with 7zip for Windows/Linux or Unarchiver for Mac.)
5. Facebook metrics
6. Bike Sharing (use `hour` data)
7. Student Performance (use just `student-por.csv` if you do not know how to merge the math grades)
8. Concrete Compressive Strength
9. SGEMM GPU kernel performance
10. Merck Molecular Activity Challenge (from Kaggle)

Merck Molecular Activity challege data. The Merck Molecular Activity challege is a much larger regression dataset than the others. The $40,000 prize went to a "deep multi-task neural network," but you do not need to evaluate that kind of model here. To download the training data, one of your group members must create a Kaggle account. Once registered, you should see the following download option:



To represent this dataset in your experiments, you should use the two training files shown above (and only those two), and use the evaluation metric of the original challenge which is described here. This metric is just the average of the squared Pearson correlation between predictions $\hat{y}_i$ and targets y_i.

The actual test data for the Merck challenge was never released, and you are not required to actually participate in the challenge itself (it is closed). So you can treat the training set like the other datasets that does not have explicit an train/test split.

The Merck features are all small integers, and each file has a different number of features. I suggest loading the initial file as follows:

```
with open(MERCK_FILE) as f:
    cols = f.readline().rstrip('\n').split(',')  # Read the header line and get list of column names

# Load the actual data, ignoring first column and using second column as targets.
X = np.loadtxt(MERCK_FILE, delimiter=',', usecols=range(2, len(cols)), skiprows=1, dtype=np.uint8)
y = np.loadtxt(MERCK_FILE, delimiter=',', usecols=[1], skiprows=1)
```

and then caching the Numpy arrays *X* and *y* to a file using np.savez so that the data can be loaded faster when ready to train a model.

Remember, if an algorithm takes more than 3 minutes to train, you have the choice of either letting the method train longer in the hopes that it will finish, or you can declare (in the report) that the method did not finish training on that dataset.

3. Classifier interpretability

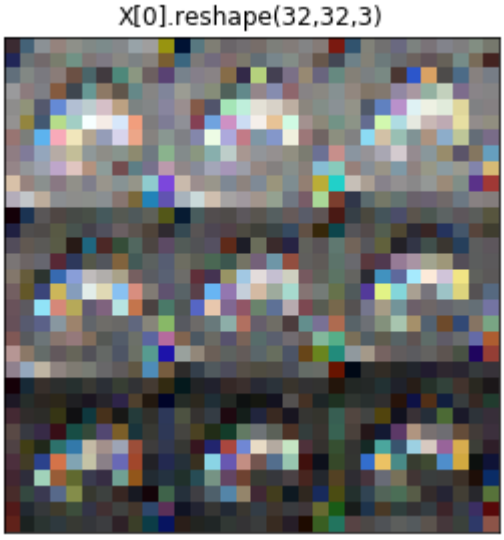
Here you will learn how to:

- load and train models on standard computer vision dataset called CIFAR-10;
- train a convolutional neural network (CNN) using PyTorch to classify images in the dataset;
- train a decision tree to classify images in the dataset;
- try to interpret the decision tree; and
- try to interpret the CNN using the 'activation maximization' technique.

Your team should report examples of what you found and state your current opinions on which of these models is most interpretable for vision data, if either of them are.

Visit the CIFAR dataset website, read about what the dataset contains, and then download the "CIFAR-10 python version." To extract the archive, use `tar -xvf cifar-10-python.tar.gz` on Linux/Mac or use `7z x cifar-10-python.tar.gz` with 7zip on Windows. Once you see the separate files (`data_batch_1` etc), follow the instructions on the CIFAR website for loading (unpickling) each data file in Python. Use the Python 3 version of the unpickling instructions. Note that to access the `data` and `labels` fields of the unpickled dictionary, you'll need to use byte keys `b'data'` and `b'labels'` rather than string keys `'data'` and `'labels'`.

Take note of the CIFAR website's description of how the images are represented. Load the first batch of images (say, into an ndarray called *X*) and try to plot the first image *X[0]*. You might think to simply reshape the length-3072 vector into shape $(32, 32, 3)$ (a 32×32 RGB image), but that would be wrong (see below).



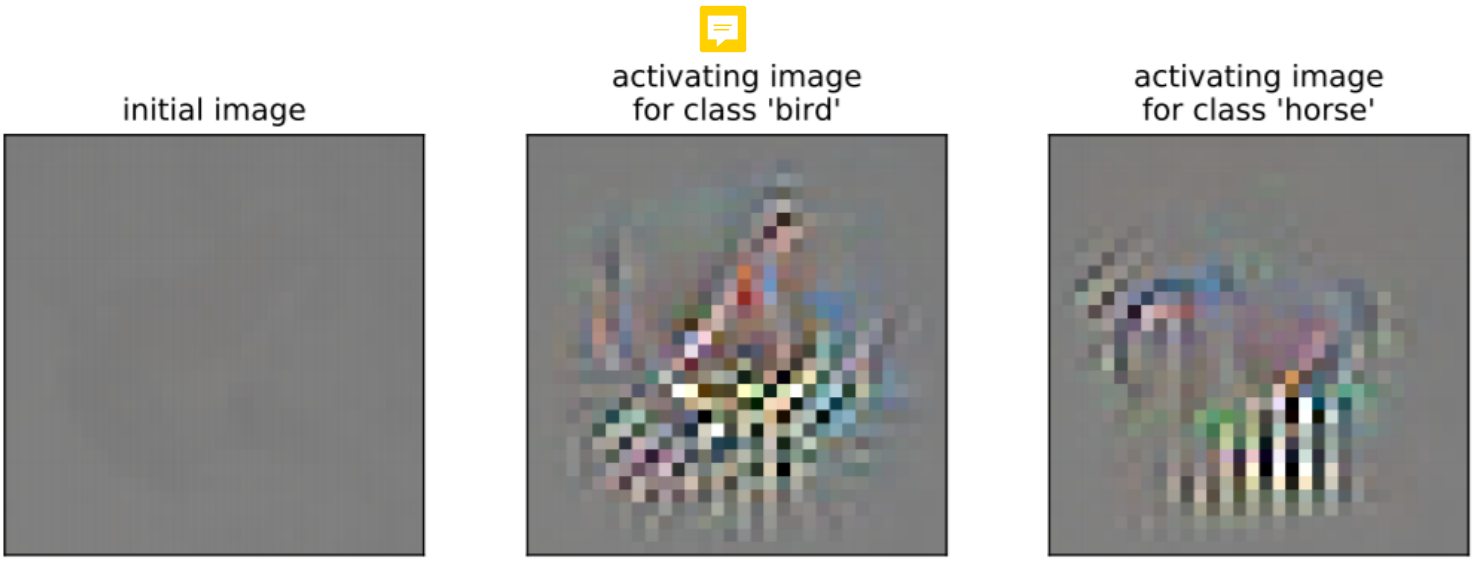
Convolutional neural networks typically expect colour image data to be in the format (N, C, H, W) which means the outermost dimension is over images $0, \ldots, N-1$, the next dimension is over colour channels $0, \ldots, C-1$ (in this case R, G, B), and the innermost dimensions are the height H and width W. So, if we want to take our (C, H, W)-formatted image and visualize it with Matplotlib (*imshow*), we need to transpose the colour axis to be the inner-most one, giving the image format (H, W, C) that Matplotlib expects. As we can see, this image probably belongs to class *frog*.

Before training a model, first write a script that combines the training batches (each of size 10,000) into a single giant batch (of size 50,000). You should end up with a 50000×3072 ndarray of dtype *np.uint8* as the features and a length-5000 ndarray *y* of dtype *np.int32* as the target labels. Likewise load the test batch and convert its targets from a `list` to an ndarray.

Decision tree classifier. Use these ndarrays to train a decision tree classifier on the entire training set. You may have to limit the tree depth for training to complete in reasonable time. Report the decision tree's accuracy on held-out test data. You can try plotting the tree like you did in lab, but use the plotting function's *max_depth* option to only plot the top few layers of the tree.

Convolutional neural network classifier. This part of the project you may not be able to complete until learning about convolutional neural networks in November. You must train a convolutional neural network on the CIFAR dataset using PyTorch. PyTorch is already installed in the lab machines, but can also be installed on your laptop. You will need to learn how to center and rescale your pixel values from $0, \ldots, 255$ of dtype *np.uint8* to be $[-1, 1]$ of dtype *np.float32*. Use any architecture you like, but keep trying until you achieve at least 75% accuracy on the test set. Train using *CrossEntropIt may take you a few attempts and each training attempt may 15-20 minutes on a laptop. If necessary, you can use techniques like data augmentation (the `torchvision` package can help with that) in order to achieve higher accuracy. Ensure that your script saves your trained model.

Activation maximization. Once your CNN is trained and saved, you can load it and start interrogating it. The idea of "activation maximization" is to ask the question: what input image would most strongly activate a particular output class prediction? Specifically, suppose we have an input image $\mathbf{x}$, which could be blank, or noise, or some other image. We feed this image through the CNN to get class activation $\hat{\mathbf{y}}$. We then choose a particular class i and ask: how should I perturb the pixels in $\mathbf{x}$ so as to increase $\hat{y}_i$? Deep learning frameworks allow us to compute the gradient of a particular class activation $\hat{y}_i$ with respect to all image pixels $\mathbf{x}$ by simply backpropagating. We can iterate this process, performing "gradient ascent" in pixel space. Below are two examples of an initial image $\mathbf{x}$ and a new image after maximizing either the 'bird' class label or the 'horse' class label. You may not see these exact results. Report what you are able to see.



To understand how this works in PyTorch, try to understand the code snippet below:

```
# Define some transformation of x, in this case a quadratic with maximum at +3
def f(x):
    return -(x - 3)**2

x = torch.tensor(0.0, requires_grad=True)
print("%.3f" % x)

for i in range(10):
    # Evaluate our function transforming x into some quantity y that we want to maximize
    y = f(x)

    # Backpropagate a gradient of y with respect to x; this fills the x.grad variable
    y.backward()

    # Move x a small step in a direction that would increase y
    x.data += 0.2*x.grad.data

    # Reset the gradient for the next iteration, since backward() always adds to the current grad value.
    x.grad.data.zero_()

    print("%.3f" % x)
```

The above code outputs the following sequence of numbers for *x*, which converge to the maximum value that $f(x)$ can take:

```
0.000
1.200
1.920
2.352
2.611
2.767
2.860
2.916
2.950
2.970
2.982
```

In activation maximization, we are doing the same thing but where $\mathbf{x}$ is an image, $f(\mathbf{x})$ is a CNN, and y is some class activation score (the value being fed into the softmax for that particular class). You must implement

activation maximization yourself. How meaningful are the patterns you see? Are you convinced that the model is learning high-level concepts like the appearance of "dog" and "cat"?

4. Novelty component

Try to introduce a novel aspect to your analysis of classifiers and regressors or to your investigation of interpretability. For example, you could add a new dataset to the study, or do a 'deep dive' one one particular dataset that's already included. You could do something as simple as try to inspect your convolutional filters as a secondary means of trying to interpret the model, and report on what patterns you saw and what they might mean. Or you could try applying interpretability techniques to one of the UCI datasets already analyzed, and report your experience.

Guidance on specific issues

Models not yet covered in lecture. Some of the models you are asked to evaluate will not be covered until later in the course, such as neural networks and Gaussian processes. That is OK, for now you can use them as a black box and take advantage of scikit-learn's unified interface (`model.fit(X,y)` , `model.predict(X)` , etc) to still apply these models to data. Or, you can wait until they're covered in the course and incorporate them into your project code. However, later on in the course you may come to understand the hyperparameters of these models better, and may want to re-run your experiments with that in mind, so making sure your experiments are automated and reproducible is important here (see "Tips on getting good marks").

Families of models. The suggested evaluation breaks models down into broad classes, such as "SVMs" rather than the more granular breakdown of "linear SVM" separate from "RBF SVM." So, we could make a statement like "SVM is the best classifier" when in fact the truth could be "RBF SVM worked best on dataset 1, whereas linear SVM worked best on dataset 2." Since these two example models belong to the SVM family, we'll count them as a win for the SVM family overall. We could likewise group models even more coarsely, as "parametric" or "non-parametric" and try to draw conclusions like "non-parametric models are the best," but this is too broad a claim to be useful. Similarly the "neural networks" family includes a broad spectrum of possible architectures (networks with many or few layers, with many or few hidden units per layer, with different activation functions, etc) and, if any one of the attempted architectures wins out on a particular data set, this individual win can be counted as a win for neural networks family as a whole. In other words, the number of layers and hidden units in a neural network can be treated as hyperparameters of "neural networks" rather than as separate types of models, and so they may need more computational investment in hyperparameter search than would a family of models that is more constrained.

Getting data sets. Data for this project comes from the UCI repository, a Kaggle challenge, and the University of Toronto CIFAR dataset website. The purpose of this project is to have you work with real data, including some minimal processing such as loading tabular data from text files. That means that, to receive full marks, you should use basic functions like `numpy.loadtxt` or load the raw data yourself using Python. If you find a specialized Python packages to fetch your data for you. You should instead download the data from its original source, either in a web browser or with `wget` , and learn how to load it and process it from that form. For example, the `torchvision` Python package comes with a CIFAR10 DataSet class that will automatically

download the raw binary data from the University of Toronto servers and serve the images as PyTorch tensors in (N, C, H, W) tensor format. Instead, you should be downloading "CIFAR-10 python version" from the CIFAR website and learn to understand and load the raw data as it was originally stored. For example, the CIFAR-10 training features are stored in five pickle files, each containing a 10000×3072 narray of dtype *np.uint8* (10000 32×32 RGB images); see the Python 3 `pickle.load` example on that website. That being said, if you are running out of time on your project and do not want to get stuck on the data loading aspect, you can try using specialized data loaders anyway and you won't lose many marks for that.

Missing values. Real-world datasets (*e.g.* from social media, businesses, hospitals) often contain "missing values" in their features. This means that the feature's actual value was unknown (unobserved). Throwing such data away is possible for the training set, but not for the test set. To still learn and to make predictions from the remaining values, one strategy is to 'impute' the missing values. The simplest way to impute a missing value is to assume (possibly incorrectly) that it probably took the most 'typical' value of that feature across the training set. This is considered a type of preprocessing, to prepare the data for training. See the scikit-learn guide for imputing missing values.

Long training times. Most of the datasets are small and training algorithms will complete in a matter of seconds. But a few datasets are of moderate size, such as the Merck and CIFAR datasets. Some models and learning algorithms do not scale to larger datasets, and so when you try to train them they will get stuck. Each time you try to train a model you should give it **at least 3 minutes** of training time, not including data loading and preprocessing. However, if the method takes longer than 3 minutes, you may decide to record that the method has "failed" to train for that particular dataset, rather than waiting to find the classification/regression performance. Models that take a flexible amount of time, such as MLPs (In real life we would give training algorithms much, much more time to complete than this, but for the sake of letting you try things.) The only exception is that when training a deep neural network on CIFAR-10 data, you should expect training time to be several minutes to an hour on a CPU, depending on the architecture and regularization you chose.

Reporting model performance. You need to propose a way of evaluating and comparing these models. Your evaluation methodology needs to be explained in the report. You can propose a methodology that you find convincing or cite an existing research paper (a benchmark) as inspiration for your chosen methodology. Your project code must ultimately evaluate every (model type, dataset) pair and then compare those individual results in such a way as to draw conclusions about which model (or models) you believe to be 'best', if any. It would help the

Using LaTeX

The report must be submitted as a PDF and written in LaTeX using the `report.tex` provided. The template is based on the submission template for *CVPR* (the IEEE International Conference on Computer Vision and Pattern Recognition. CVPR is the top conference in the field of computer vision and machine learning. Its two-column format is particularly compact.

The vast majority of research papers in the engineering and mathematical sciences are written in *LaTeX*. LaTeX is a "compiled" document format, unlike for example Micorosft Word. If you do not know LaTeX, you can learn from the internet. Beginners should start drafting their report early, even if results are not ready, because you

will for sure have LaTeX formatting questions or problems; you don't want to be stuck on formatting the night before the project is due.

In terms of software tools, you have multiple options:

- **Overleaf.** This is an online collaborative LaTeX document authoring service. The free version only allows for one official `collaborator' per document, but the creator of the document can still share it by link with others. That way, several people can still edit and preview the document online. Nothing to install, but you need to upload your figures to their system. See overleaf.com.
- **Visual Studio Code.** This is an integrated development environment (IDE) that runs locally on your computer and supports many types of files (including `.py` and `.tex`) through extensions that are very easy to add. The LaTeX extension supports PDF preview. If you go this route, you'll need to find a way for multiple people in your group to collaborate on `report.tex`, such as Dropbox or Github. See code.visualstudio.com and add the LaTeX Workshop extension.
- **MikTeX.** This provides a set of local command-line tools for compiling `.tex` files to other formats like `.pdf`. It comes with a pre-packaged LaTeX text editor with PDF preview. If you go this route, you'll need to find a way for multiple people in your group to collaborate on `report.tex`, such as Dropbox or Github. See miktex.org

Citing books or papers in your report

The LaTeX report template comes with a bibliography and examples of how to cite it. You should look at the UCI page for each data set used in the project and, if there is an explicit "Citation Request" for a particular paper, you should **cite that paper** when you're listing the data sets that you use. (If there are multiple papers listed, pick one.) If the UCI page merely mentiones "related papers", you do not need to cite them.

For example, your report might include LaTeX text like:

```
We used the following classification datasets: diabetic retinopathy~\cite{antal2014ensemble}, ...
```

where you have already added the corresponding item to the `bibliography.bib` file as part of your report:

```
@article{antal2014ensemble,
  title={An ensemble-based system for automatic screening of diabetic retinopathy},
  author={Antal, B{\'a}lint and Hajdu, Andr{\'a}s},
  journal={Knowledge-based systems},
  volume={60},
  pages={20--27},
  year={2014},
  publisher={Elsevier}
}
```

This will be rendered in the document as something like "diabetic retinopathy [5]" and the paper details for *Antal et al* will be added to your bibliography the next time you recompile your LaTeX document.

A good way to quickly find the "bibtex entry" for a particular paper is to copy its title, paste the title into Google Scholar, find the paper in the search results, and click the little blue quotes icon that looks like this:

This will give you a dialog with a link containing a "BibTeX" entry for the paper you're searching for. Paste the BibTeX entry into your `bibliography.bib` file, cite it in the text, and recompile to see the new entry.

If your report refers to other papers or books, please add them as citations like the above. However, if your 'novelty component' involves applying a technique that you found in another paper, you should include a citation to that paper. If the technique is based on a blog post, you should instead include a URL to that blog post, either in the bibliography or using the `\url{...}` provided by the `hyperref` package, which is already included in the template.

Tips for getting a good grade

- **Navigating uncertainty.** The project is not a checklist ("if you do X,Y,Z then you're done"). This leads to uncertainty over what *precisely* is expected, and this uncertainty can understandably lead to anxiety. Try not to worry. Uncertainty and anxiety is normal part of navigating a project with open-ended aspects. Your instructor and TAs will try to guide you with specific questions that you may have, but the questions have to come from you. Use the Moodle discussion board for general project questions and Moodle direct messages for group-specific questions.
- **Apply practices from lecture.** The general idea is to apply the techniques you learned in lecture. Try to be systematic in your analysis, for example ensuring the same hyperparameter search and training procedures are applied to all data sets. Obviously something like "training on the testing data" is a huge no-no and would render your conclusions invalid. You are of course free to explore other techniques if they interest you or if you believe that they will provide insight or surprises to the analysis.
- **Aim for reasonably-good predictive accuracy.** If you report that "model type *X* is the best because it it had an average test performance of 84% across the data sets" and yet most other groups found that model type *X* got an average test performance of 98%, then this may reflect some problem in the way that you either processed the data or trained models of type *X*. However, you will **not lose marks** just because a few other groups got better prediction accuracy than you on a particular dataset, or even just because some other group came to a different conclusion about what was "best." As long as your predictive accuracy is not indicative of a bug or weak attempt at modeling, do not worry about fine-tuning performance.
- **Reproducible results.** Always set the random seed or random state of whatever Python framework you're about to rely on, be it numpy, sklearn, or torch. Don't end up with a different train/test split the next time you run your training pipeline! At least not unintentionally!
- **Automate end-to-end as much as possible.** The more that your project code resembles an automated pipeline that goes from raw downloaded data to performance numbers, the better. In the ideal, it should be possible to generate *all* your results by running a single master script. This script would trigger individual scripts that run more specific groups of experiments (load training data, preprocess, hyperparameter search and final training, save the model, load the model, load the test data, generate raw performance numbers, dump the raw performance numbers to one or more files). If a single master script is not possible, that's OK, but keep them to a minimum. The final performance numbers that you report in your tables and text should ideally come from running scripts, and *not* be the result of some transient IPython notebook code cells for which you have manually been running indvidual code cells. You can rely on IPython notebooks to help you make plots (PNG or PDF) from that raw data, but the Jupyter notebook should not

be generating the raw results. Use Jupyter notebooks primarily for prototyping, if you wish. You should also try to avoid manually processing the raw data, such as editing the raw data files in a text editor; if a particular data set (say, a UCI data set) comes in an odd format that needs custom processing (removing header lines, converting categorical strings to numbers, etc) then that custom processing should ideally be done by a Python script, not directly by your fingers! You can of course cache intermediate results to make it faster to initiate training, such as storing a Numpy representation of the cleaned-up data. Still, if you are too pressed for time to be this systematic, you can still simplify the raw files "by hand" for the sake of expedience and you will not lose many marks for that.

- **Quality of scientific presentation** is important for this course. That means presenting figures, plots, and tables that are maximally informative for the scientific question being asked. For example, students should know when to use a scatter plot, a bar plot, a box plot, etc. Employers and professors often find that students and recent graduates lack these skills, so this is the best time to learn as much as possible! Use examples of plots and tables from labs or from your favourite research papers as a guide.
- **Quality of Python code** is minor for this course but will still have a small impact on grade. Employers and professors find Python coding one of the most important skills that ML students and graduates lack. For this course, goal is not to punish inexperienced coders, but rather to encourage students to learn as much Python as they can during this course. Students working in ML need good Python coding skills, so try to learn from the best examples, and if you're in doubt about your code quality ask a TA for his/her opinion, so that you can take this opportunity to learn.
- **Quality of writing** is somewhat important for this course. Apart from Python coding, scientific writing is the next thing that employers and professors wish that students and graduates were better at. That being said, this is a machine learning course, not a writing course, so poor writing will not ruin your grade, and minor grammatical errors will have no impact. But you are unlikely to get top grades if you have a report that is unclear or incoherent, even if your code itself is clean, concise, and systematic.
- **Methodological and code consistency across the project.** For the predictive accuracy component, avoid submitting a project where it's clear that each team member worked independently and did not share a common underlying methodology or codebase. Here are some examples where a team missed opportunities to consolidate their code/ideas:
  - One team member does classification, the other does regression, and they each use completely different methods and/or code for doing hyperparameter search.
  - Multiple versions of code to load data sets that are in the same format. If two data sets are in the same or similar format, they should be loaded by the same code.
  - The project code contains several scripts that must be run to reproduce results, but each script follows a different style for accepting arguments (input directories, output directors, etc.)
  - Different versions of code that generate what is essentially the same kind of plot.