DS4023 Machine Learning

Lecture 1:
Introduction to Machine Learning

Mathematical Sciences
United International College

Outline

- Introduction to machine learning
- Machine learning models and tasks
- Model evaluation and model selection

What is Machine Learning

- Learning = Improving with experience at some task
  - Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more effectively the next time

What is Machine Learning

- Definition:
  - Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.
  - Tom Mitchell (1998) Well-posed Learning Problem: A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

What is Machine Learning

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?

1. Classifying emails as spam or not spam.
2. Watching you label emails as spam or not spam.
3. The number (or fraction) of emails correctly classified as spam/not spam.
4. None of the above—this is not a machine learning problem.

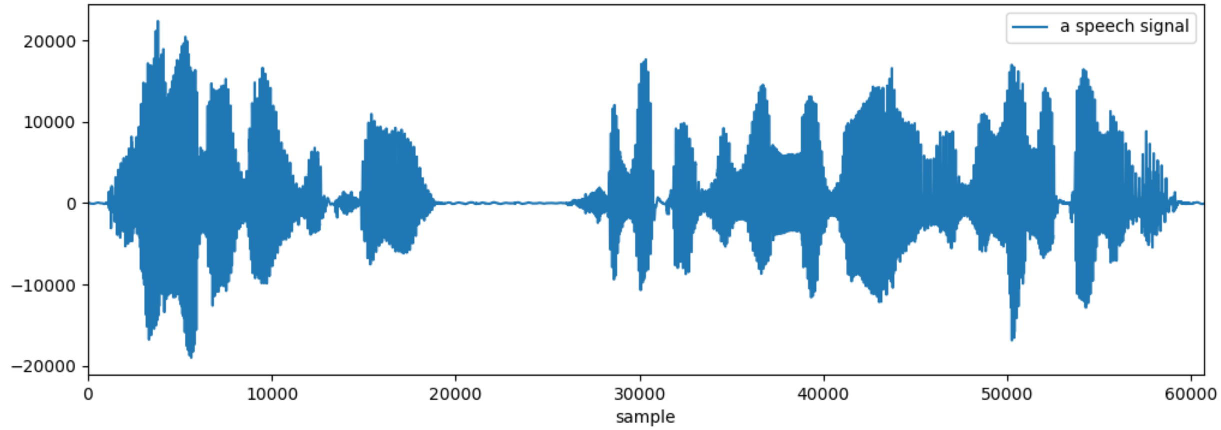
Machine Learning is Everywhere

- Character recognition
  - raw data: image

  - class: numerals, English (Chinese, etc.) characters

Machine Learning is Everywhere

- Speech recognition
  - raw data: speech signal



  - class: spoken words

Machine Learning is Everywhere

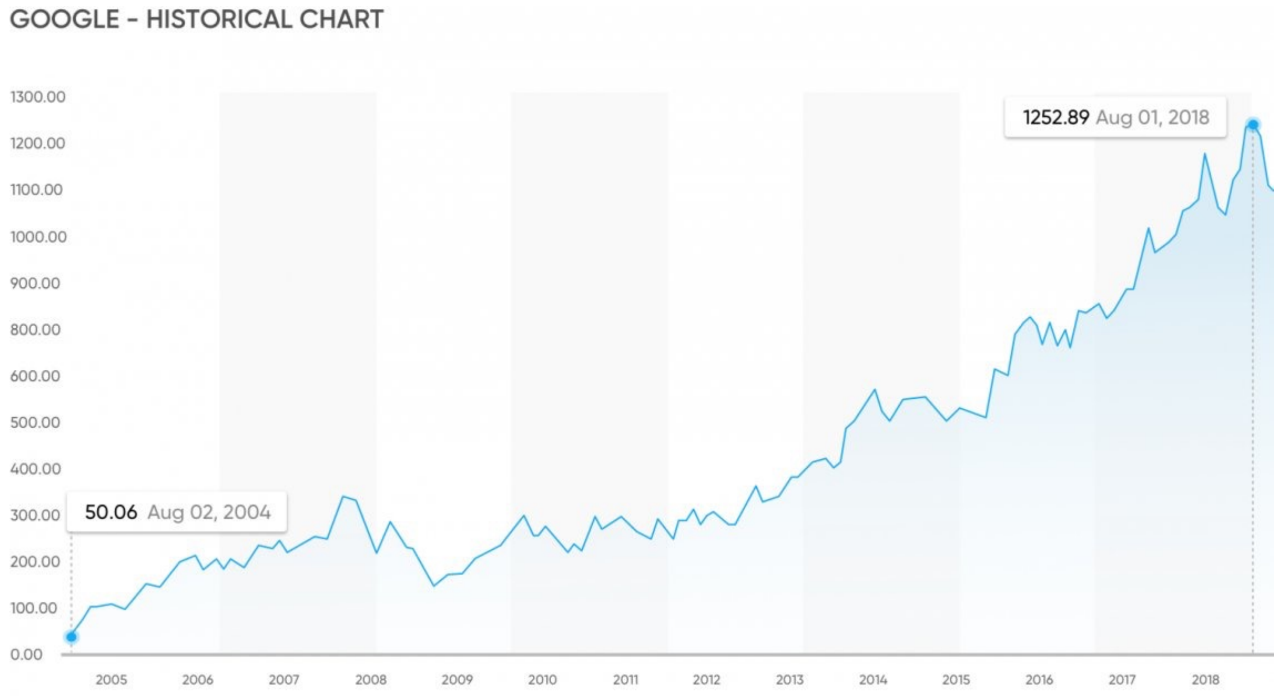
- Document classification
  - raw data: (web) document

> *As the movie year winds down, I would like to express my gratitude to Martin Scorsese. Not only for making "The Irishman," his best movie in a long time and one of the best of 2019 (see below), but also for reminding the world of the value of cinema.*
> *The art form is in one of its periodic identity crises. A big chunk of our collective attention — we don't yet know how big, or with what consequences — is migrating to streaming platforms whose offerings include a lot of the stand-alone single-episode narratives that we used to see mainly in theaters.*

  - class: semantic categories (movie, art, money,...)

Machine Learning is Everywhere

- Financial engineering
  – raw data: financial time series (e.g., stock prices)



  – task: classify financially healthy / unhealthy company, stock prediction, etc.

Machine Learning is Everywhere

- Many many other examples
  - Facebook: photo tagging, ranking articles to your news feed
  - Amazon: eCommerce fraud detection, forecasting demand, pricing
  - NASA: identifying stars, supernovae, clusters, galaxies, quasars, exoplanets, etc.
  - Google Spreadsheets: uses machine learning to fill in missing values

Learning Paradigm

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Supervised Learning

- The learner is provided with a set of inputs together with the corresponding desired outputs
  – Has a “teacher”
- Example:
  – teaching kids to recognize different animals
  – graded examinations with correct answer provided

Supervised Learning

- Breast cancer (malignant/benign classification)
  - Input: Tumor size samples
  - Output: Whether the tumor is malignant or benign
  - Task: Learn a classifier from the provided input and output, that can predict the label for new tumor size inputs.

Supervised Learning

- House price prediction (Regression)
  - Input: House size samples
  - Output: House prices
  - Task: Learn a model from the provided input and output, that can predict the house prices (quantity) for new house size inputs.

Supervised Learning

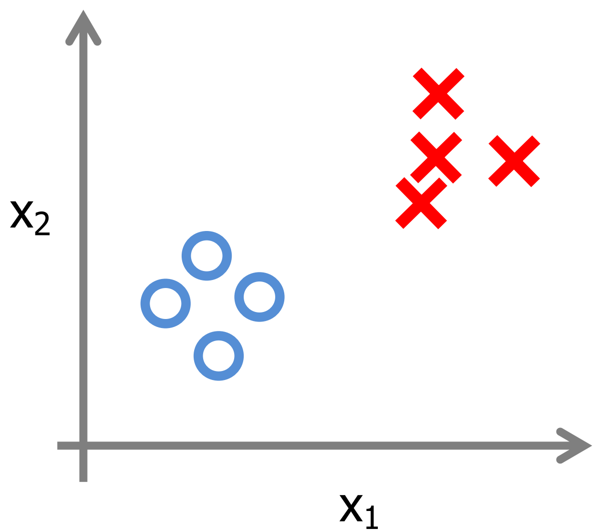
- You’re running a company, and you want to develop learning algorithms to address each of two problems.

- Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
- Problem 2: You’d like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

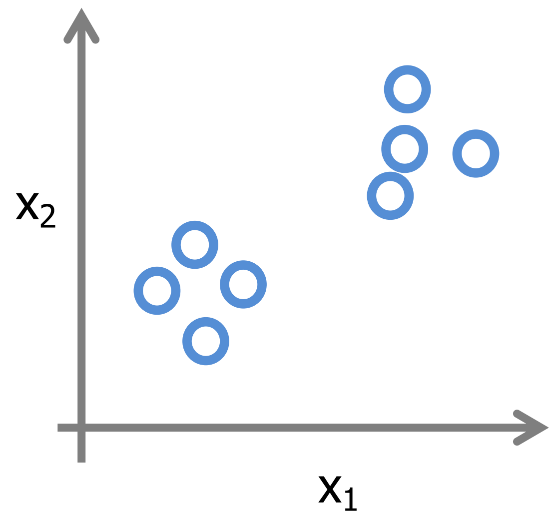
- Should you treat these as classification or as regression problems?

Unsupervised Learning

- Training examples as input patterns, with no associated output
  - no “teacher”
  - similarity measure exists to detect groupings/ clusterings



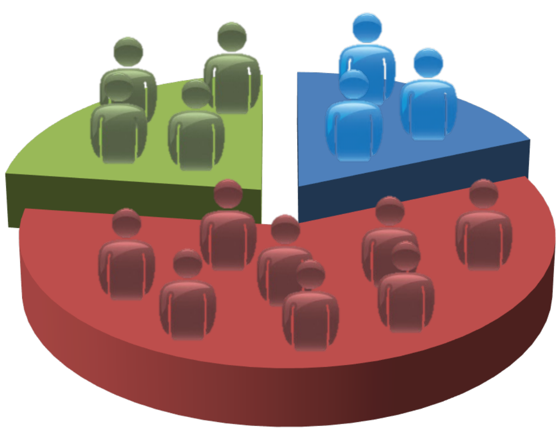
Supervised Learning



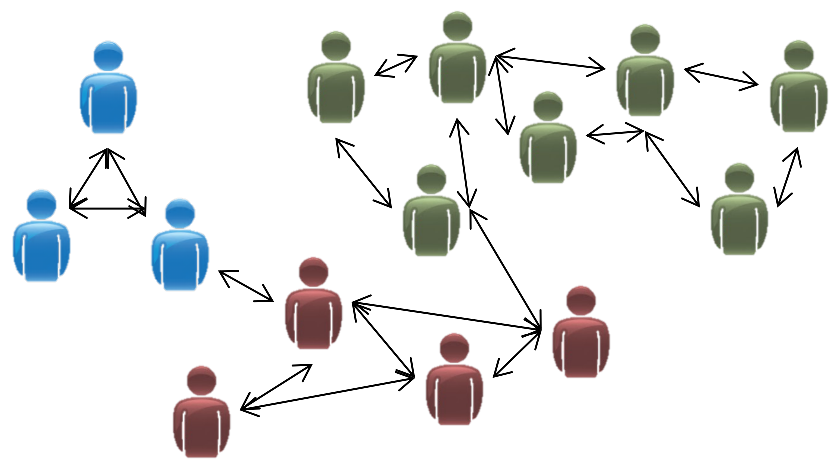
Unsupervised Learning

Unsupervised Learning

- Clustering
  - in the early stages of an investigation, it may be helpful to perform exploratory data analysis to gain some insight into the nature or structure of the data



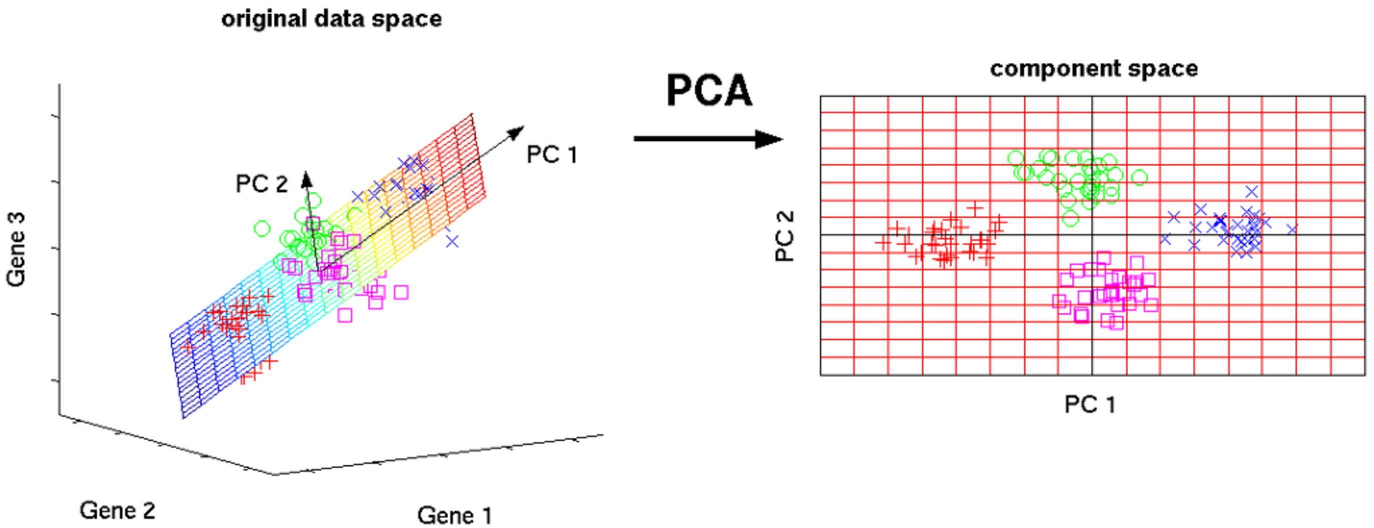
Market segmentation



Social network analysis

Unsupervised Learning

- Find features or preprocess existing features for the subsequent pattern classification problem (supervised learning)
  – Principle component analysis (PCA)

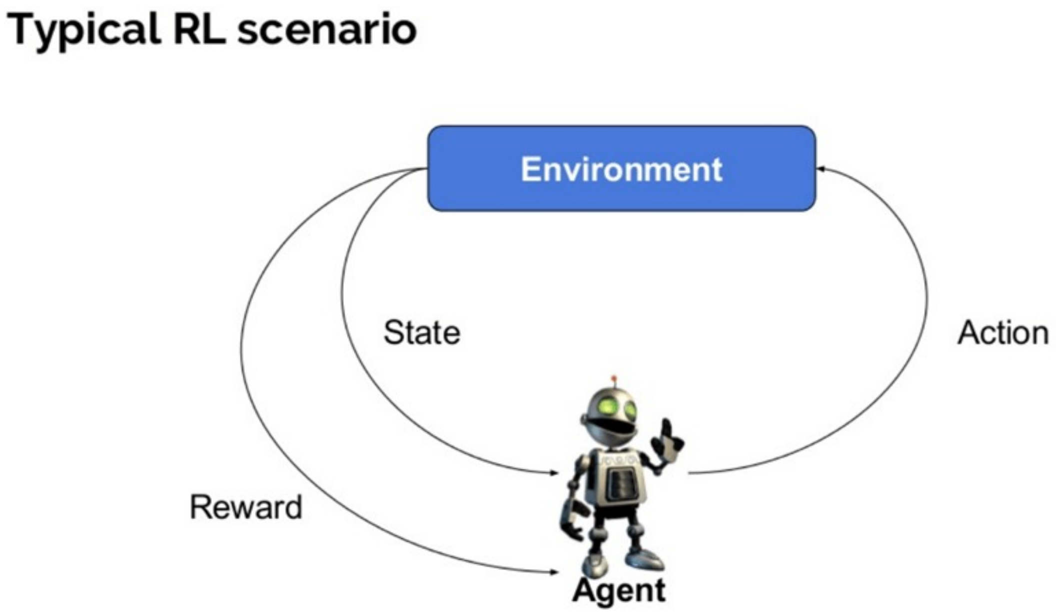


Reinforcement Learning (RL)

- Training examples as input-output pairs, with evaluative output only
  - try to increase the reinforcement it receives
- Example
  - Graded examinations with only overall scores but no correct answers.

Reinforcement Learning

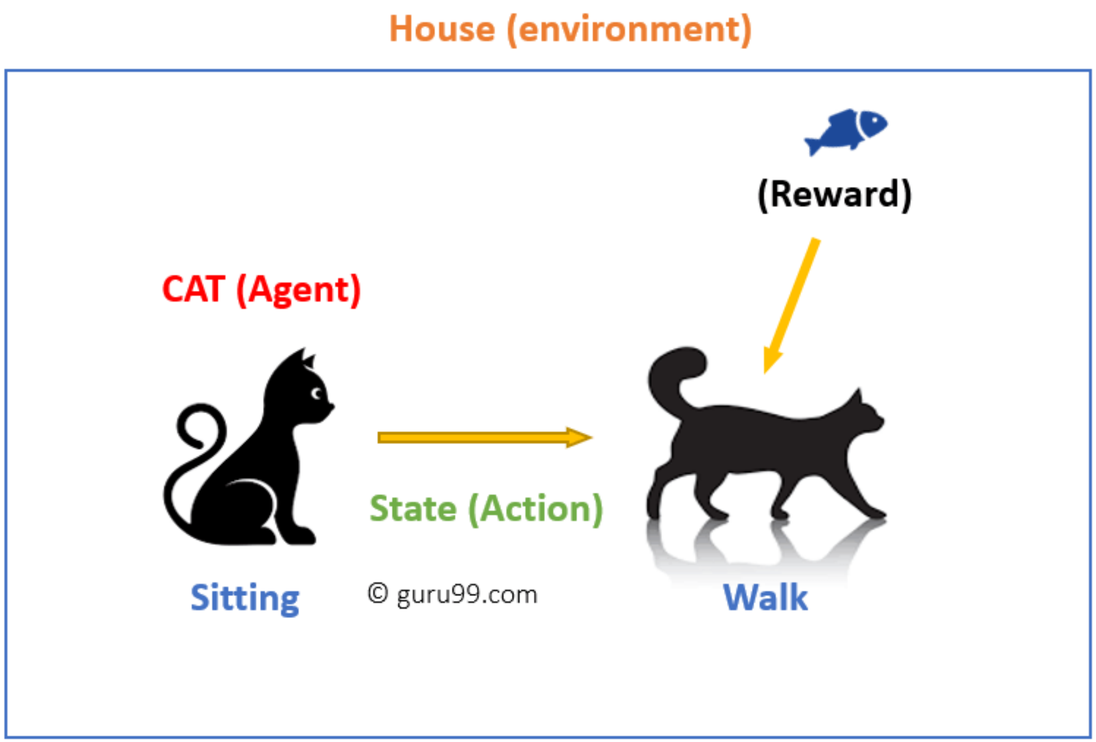
- Learning from interacting with an environment to achieve a goal
  - Learning a mapping from states to actions to maximize long-term reward



Reinforcement Learning

- Consider the scenario of teaching new tricks to your cat
  - As cat doesn’t understand human language, we can’t tell him directly what to do. Instead, we follow a different strategy.
  - We emulate a situation, and the cat tries to respond in many different ways. If the cat’s response is the desired way, we will give him fish.
  - Now whenever the cat is exposed to the same situation, the cat executes a similar action with even more enthusiastically in expectation of getting more reward(food).
  - That's like learning that cat gets from "what to do" from positive experiences.
  - At the same time, the cat also learns what not do when faced with negative experiences.

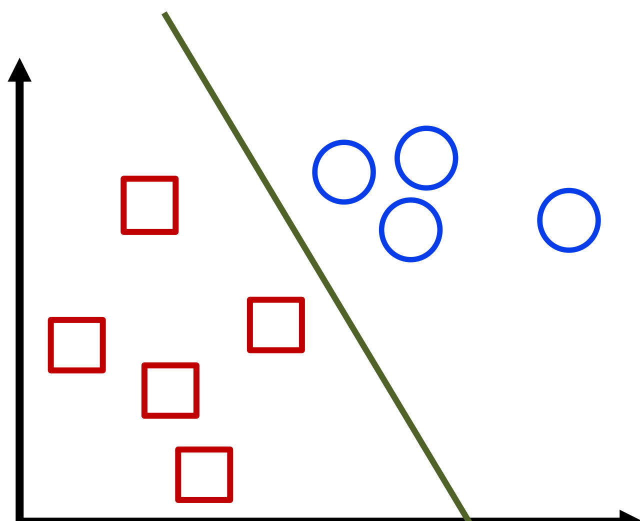
Reinforcement Learning



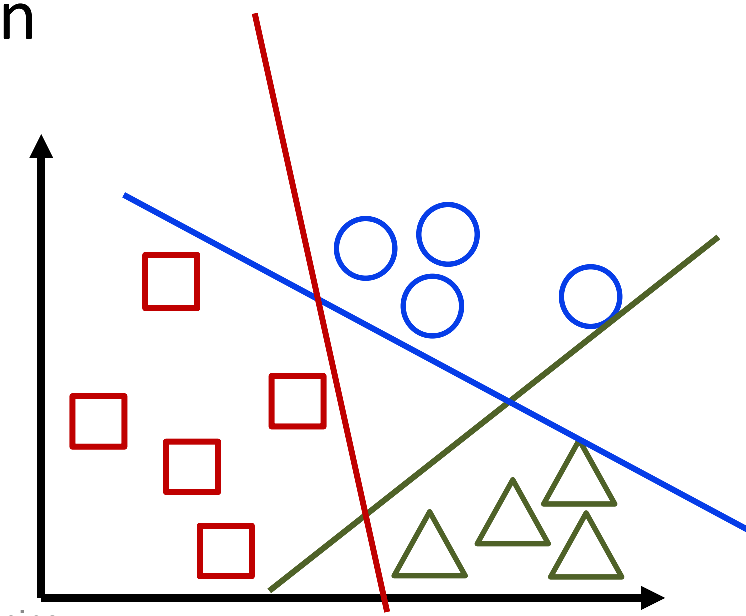
- Goal: Maximize reward
- State: Sitting, Walk
- Action: Transition from one state to another
- Reward: Fish

Variants and Extensions

- Multiclass classification is the problem of classifying instances into one of three or more classes.
  – Classifying instances into one of two classes is called binary classification



Binary Classification



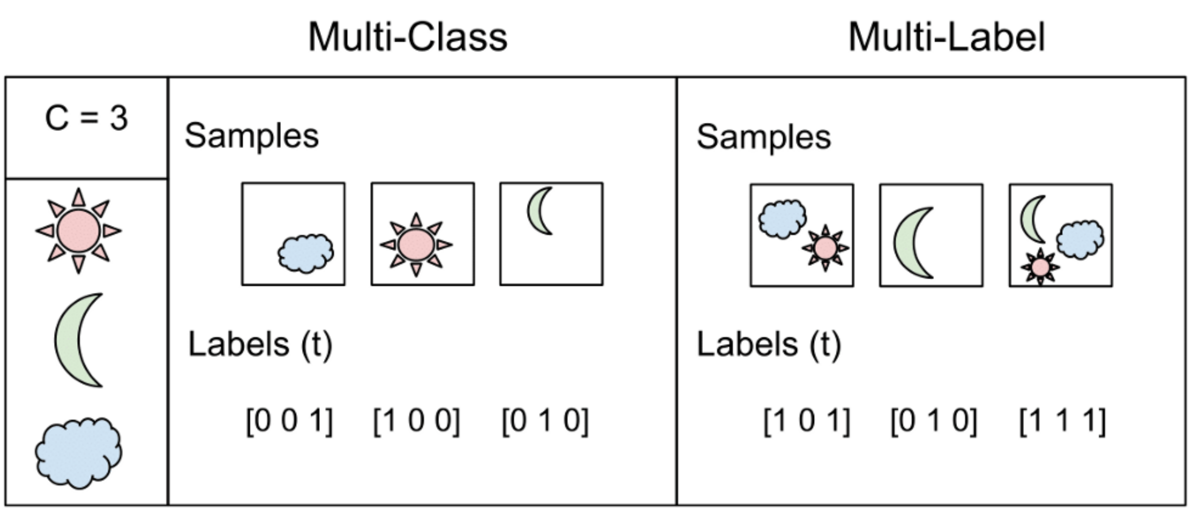
Multiclass Classification

Multiclass Classification

- Reducing multiclass classification to binary classification
  - one-vs-rest: 1. binary classification “C_i” vs “not C_i”; 2. select the class that is most certain
  - one-vs-one: 1. binary classification “C_i” vs “C_j”; 2. aggregate the results by voting
  - How many classifiers you need to train for each of the strategy?

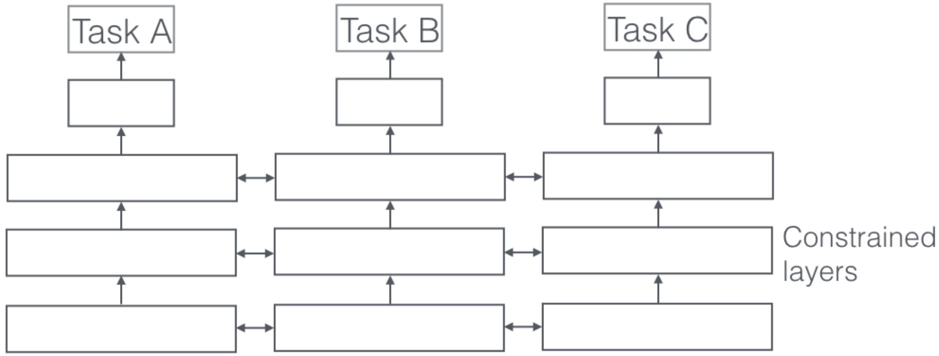
Variants and Extensions

- Multilabel classification is the single-label problem of categorizing instances into equal/more than two classes
  - An instance can have multiple labels, e.g., image tagging, text categorization,...
  - Transform into binary classification or multiclass classification problem...



Variants and Extensions

- Multi-task learning (MTL) is a subfield of machine learning in which multiple learning tasks are solved at the same time, while exploiting commonalities and differences across tasks.
  - This can result in improved learning efficiency and prediction accuracy for the task-specific models, compared to training the models separately.



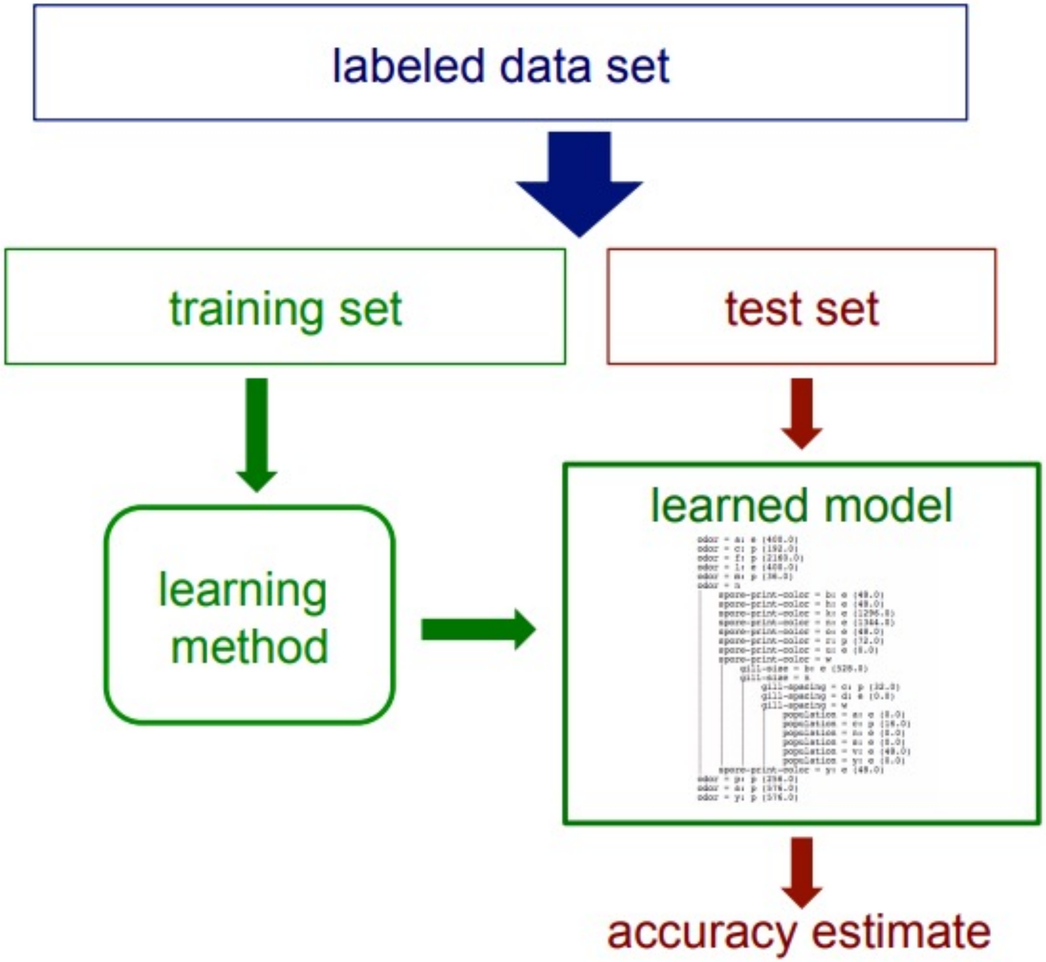
Model Evaluation and Selection

- Generally, the difference between the actual predicted output of the learner and the true output of the sample is called "error"
  - Training error/ empirical error: the error of the learner on the training data
  - Generalization error: the error on the new data

Model Evaluation and Selection

- We want to get a learner with a small generalization error
- However, we do not have the information for new data, instead we try to minimize the empirical error on the training data
  - split data randomly into a training set and a test set (e.g., a 70%/30% split).
  - train your model on the training set and see how it performs on the test set.
  - use the "testing error" on the test set as an approximation of the generalization error

Model Evaluation and Selection



Model Evaluation and Selection

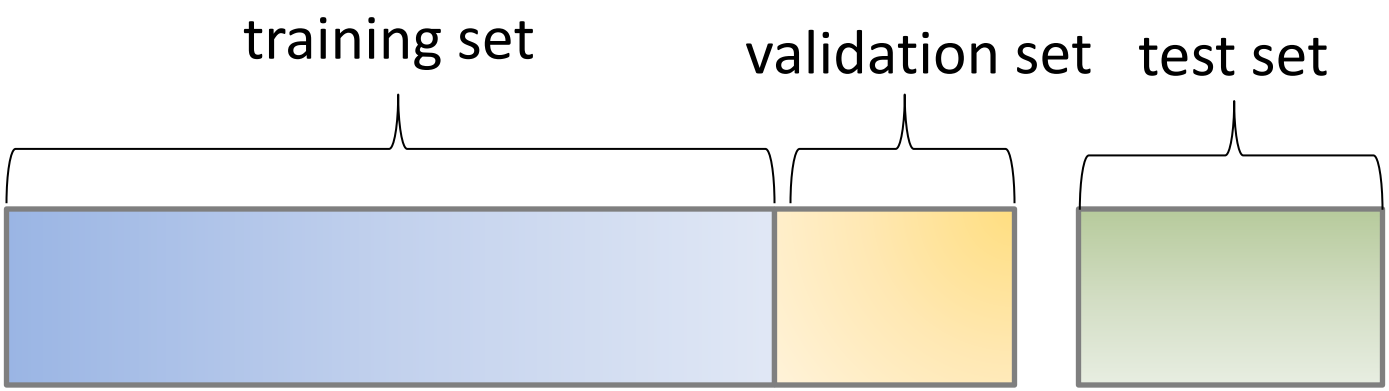
- Strategies for generating training and testing datasets
  - Hold-out: just set aside some portion of the data for testing
  - Cross validation:
    - partition data into k disjoint datasets (called folds) of approximately equal size; iteratively take k-1 folds for training and validate on the remaining fold ; average the results
  - Boostrapping:
    - new datasets are generated by sampling with replacement (uniformly at random) from the original dataset; then train on the bootstrapped dataset and validate on the unselected data

Model Evaluation and Selection

- Learning algorithms and models have hyperparameters, whose values are set before learning
  - k value in k-nearest neighborhood
  - learning rate for a neural network
- Hyperparameter tuning is choosing a set of optimal hyperparameters for learning algorithm
  - First split the dataset into training and test.
  - Keep aside the test set, and randomly choose x% of training set to be the actual trainining set and the remaining $(100 - x)$% to be the validation set.

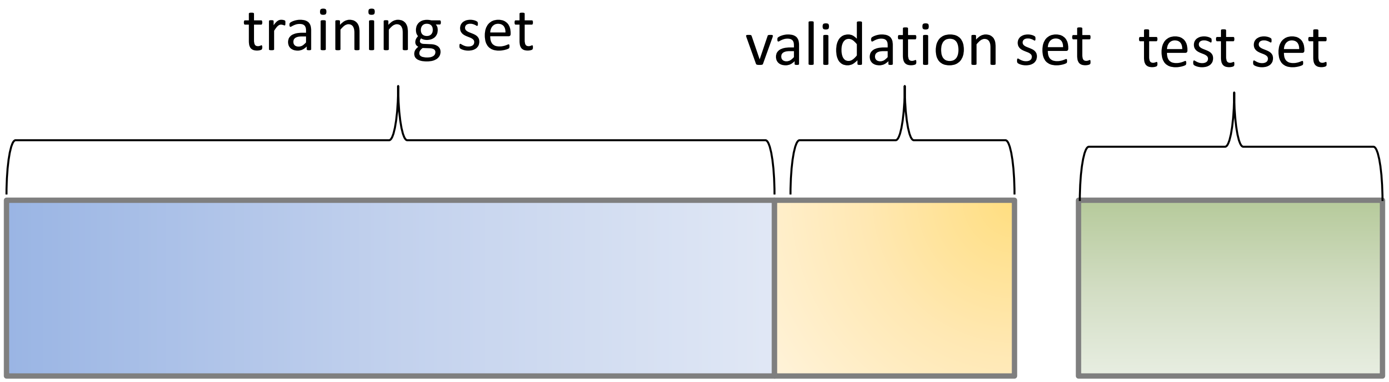
Model Evaluation and Selection

- Data samples
  - Training set: The sample of data used to fit the model.
  - Test data: The sample of data encountered in actual use of evaluating the completely trained/learned model.
  - Validation set: The sample of data used in model evaluation and selection while tuning model hyperparameters.



Model Evaluation and Selection

- Hyperparameter tuning methods
  - Grid search: works by searching exhaustively through a specified subset of hyperparameters.
  - Random search: differs from grid search mainly in that it searches the specified subset of hyperparameters randomly instead of exhaustively. The major benefit being decreased processing time.



Evaluating Regression Models

- Given dataset $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$
- Residue: a residue e_i is the difference between the observed and predicted outcome:

$$e_i = y_i - f(x_i)$$

- Mean squared error (MSE)

$$E(f; D) = \frac{1}{m}\sum_{i=1}^{m} {e_i}^2$$

- Coefficient of determination or R-squared

$$R^2 = 1 - \frac{\sum_{i=1}^{m} {e_i}^2}{\sum_{i=1}^{m} (y_i - \bar{y})^2}$$

Evaluating Classification Models

Confusion Matrix

Predicted class

Actual class

Classes	C_1	C_2	Total
C_1	true positives (TP)	false negatives (FN)	positives
C_2	false positives (FP)	true negatives (TN)	negatives

For binary classification problem:

- Precision = $\frac{TP}{TP+FP}$
  - Fraction of relevant instances among the retrieved instances
- Recall = $\frac{TP}{TP+FN}$
  - Fraction of relevant instances that have been retrieved over the total amount of relevant instances
- F-measure = $2 \cdot \frac{precision \times recall}{precision + recall}$
  - combines precision and recall (harmonic mean)

Evaluating Classification Models

- In some scenario, we may have multiple confusion matrices
  - perform multiple training/test
  - training/test on multiple datasets
  - multiclass classification
- Macro average:
  - get the precision and recall of each confusion matrix, $(P_1, R_1), (P_2, R_2), \ldots, (P_\mathrm{n}, R_\mathrm{n})$
  - $macro\text{-}P = \frac{1}{n}\sum_{i=1}^{n} P_i$
  - $macro\text{-}R = \frac{1}{n}\sum_{i=1}^{n} R_i$
  - $macro\text{-}F1 = 2 \cdot \frac{macro-P \times macro-R}{macro-P + macro-R}$

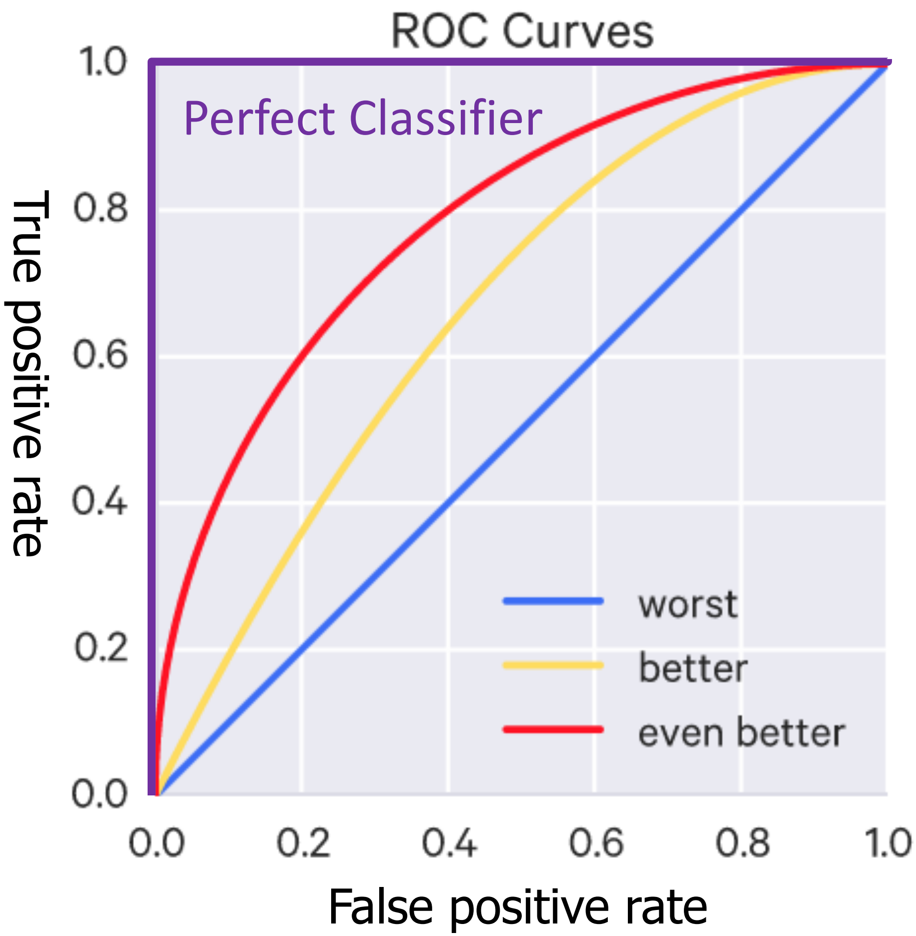
Evaluating Classification Models

- In some scenario, we may have multiple confusion matrices
  - perform multiple training/test
  - training/test on multiple datasets
  - multiclass classification
- Micro average:
  - get the average of each item in the confusion matrices $\overline{TP}$, $\overline{TN}, \overline{FP}, \overline{FN}$
  - $micro\text{-}P = \frac{\overline{TP}}{\overline{TP}+\overline{FP}}$
  - $micro\text{-}R = \frac{\overline{TP}}{\overline{TP}+\overline{FN}}$
  - $micro\text{-}F1 = 2 \cdot \frac{micro-P \times micro-R}{micro-P + micro-R}$

Evaluating Classification Models

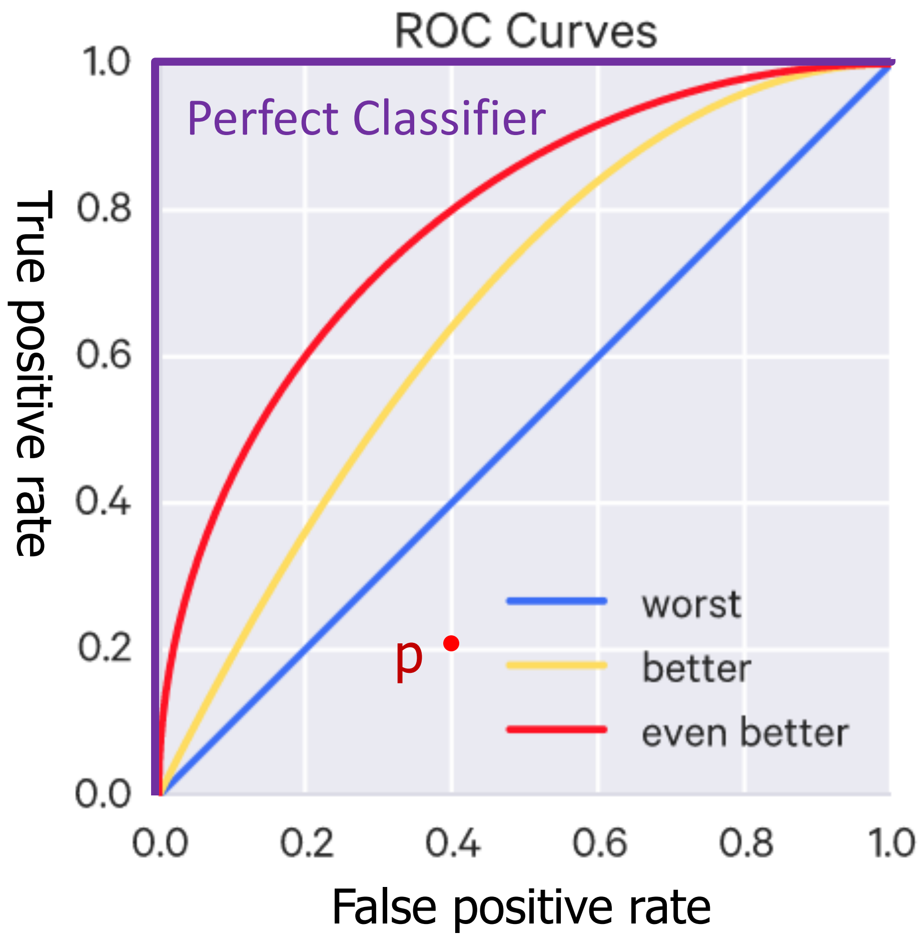
- The classifier or diagnosis result can be an arbitrary real value (continuous output), the classifier boundary between classes must be determined by a threshold value
- A receiver operating characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied
  - The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings
  - True positive rate: $TPR = TP/(TP+FN)$)
    - Recall or sensitivity
    - How often are positive-labeled samples predicted as positive?
  - False positive rate: $FPR = FP/(TN+FP)$)
    - How often are negative-labeled samples predicted as positive?

Evaluating Classification Models



- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. (Random guess)
- In real, the ROC curves are not smooth, jagged plot.

Evaluating Classification Models



- Effect of threshold on ROC curve?
- Plot the jagged ROC curve?
- What if we have a point lies under the diagonal curve (point p)?

Evaluating Classification Models

- Area under ROC curve (AUC) is used to quantify how good the classification algorithm
  - In general, an AUC of above 0.8 is considered "good"