Evaluation

- **Performance vs Effectiveness vs Efficiency**
  - Effectiveness: doing the right things
  - Efficiency: doing things right
    - **Achievement per unit of input**
  - Performance: a synonym for effectiveness? An umbrella term?
- **The question: To what extent is a machine-learning algorithm achieving its objective?**
  - What is the objective?
  - How can it be measured?
  - Hard to answer… e.g. if the objective is to maximize economic strength… what does that mean? How do you measure it?
- **The Standard Objective and Evaluation Types**
  - The objective of the world of research and teaching **- make classifications and predictions as well as possible**
    - Minimize errors, maximize e.g. precision, accuracy, etc..
  - **Costs**
    - Rarely considered in academia
  - **Offline vs Online evaluation**
    - **Offline**
      - Measures success on historical data
      - RMSE, accuracy, precision etc
      - **Offline evaluation metrics**
        - **Classification**
          - Confusion matrix

		Actual Class	
		True Positive	**True Negative**
Predicted Class	**Predicted Positive**	True Positives (TP)	False Positive (FP)
	Predicted Negative	False Negative (FN)	True Negatives (TN)

          -
          - **E.g.** in disease detection, the detection rate is common (TP rate) **but we need more**
            - **TP has little meaning alone**
            - **Can always be 100% if you compromise on FP**
            - In cancer screening misclassification is high cost
              - FN is huge, FP is minimal
          - **Accuracy: correct predictions/all predictions (micro)**
            - Not necessarily meaningful
            - Can use a heuristic to predict modal class and accuracy is high

        - Accuracy can guide improvements - calculate accuracy, and analyse failures, add heuristics
- **Average pre-class accuracy**
    - **Macro**
    - Smoothens outlier classes
    - Not used in real world
    - $$AveragePerClassAccuracy = \frac{\sum_{i=1}^{n} ClassAccuracy_i}{n}$$
- **Log-Loss**
    - **"Soft" measure for accuracy for probabilistic classifiers**
    - Considers the distance to correctness
    - **Cross entropy between the distribution of the true labels and the predictions**
        - **Entropy measures unpredictability**
        - **Cross entropy incorporates the entropy of the true distribution** plus the **extra unpredictability** when one assumes a **different distribution than the true dist.**
    - Log-loss: information-theoretic measure to gauge the "extra noise" from using a predictor as opposed to true labels
    - **Equal weight for FP and FN**

$$LogLoss_{MultiClass} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(p_{i,j})$$

N = number of instances
M = number of classes, i.e. labels
$p_{i,j}$ = binary indicator if label j is correctly predicted for instance i [0/1]
$p_{i,j}$ = probability that j is the correct label for i

    - $$LogLoss_{BinClass} = -\frac{1}{N}\sum_{i=1}^{N} [\log p_i + (1 - y_i)\log(1 - p_i)]$$
- **Receiver Operating Characteristic Curve (ROC Curve)**
    - TP vs FP plot
    - **Shows how many TP can be gained by accepting more FP**
    - **Difficult to see which algorithm is better => Area under the curve (auc)**
    - Values between 0 to 1 (in practice 0.5 to 1)
    - AUC = 0.5 is random classification

- **Ranked retrieval metrics (Classification)**
    - **Precision**
        - P for positivity
            - TP/(TP+FP)
        - **p@n is the precision among the top n results for search engines**
    - **Mean Reciprocal Rank**
        - Measures at which rank the first relevant result is displayed
        - **Reciprocal Rank of the first relevant result**
        - **The average of the reciprocal ranks**
        - MRR only cares about the first relevant result

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

        - $rank_i = rank\ of\ first\ relevant\ result\ for\ query\ i$
          $|Q| = numberof\ search\ queries$
    - **Mean Average Precision (MAP)**
        - **Average precision** (for one query)

$$\mathrm{AP}(Q_i) = \frac{1}{|R|} \sum_{j=1}^{|R|} p@k$$

            - Q_i = the i-*th* query
              $|R| = number\ of\ relevant\ results$
              $R = ranks\ of\ relevant\ results$
              $p@k = precision\ at\ rank\ k\ (k \in R)$
        - **Mean average precision (over all queries)**

$$\mathrm{MAP} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} AP(Q_i)$$

            - $|Q| = numberof\ search\ queries$
    - **Normalized Discounted Cumulative Gain (nDCG)**
        - idea : more relevant items should be ranked higher than less relevant items
        - **Cumulative Gain:** sum of top k items' relevance (not accounting for position)
            - $CG_k = \sum_{i=1}^{k} rel_i$ $\quad rel_i = relevance\ at\ postion\ i$

- **Discounted Cumulative Gain:** discounts relevant items which are ranked too low
  - $DCG_k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}$ $\quad Alt.DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}$
- **Normalized DCG:** DCG normalized to be between 0 and 1
  - $nDCG_k = \frac{DCG_k}{IDCG_k}$ $\quad IDCG_k = \sum_{i=1}^{|REL|} \frac{2^{rel_i} - 1}{\log_2(i+1)}$

- **Recall**
  - How many of the relevant items were retrieved?
    - Goal is to find as many relevant docs as possible
    - bot not concerned with irrelevant docs in the results
  - TP/(TP+FN)
- **Precision-Recall Curve**
  - Plot of precision vs recall
  - **F-Measure/F_1 combines precision and recall into one metric (harmonic mean)**
  - Precision-Recall curve & F-Measure is similar to ROC curve and AUC
  - $$F_\beta = (\beta^2 + 1) \times \frac{P \times R}{(\beta^2 \times P) + R}$$
    $\beta < 1$ emphasizes precision
    $\beta > 1$ emphasizes recall
  - $$F_{\beta=1} = F_1 = 2 \times \frac{P \times R}{P + R}$$

- **Regression Metrics**
  - Can typically evaluate as classification/ranking
  - **E.g. movie rating prediction (1-5 stars)**
    - Treat as a ranked list/classification
  - **Mean Absolute Error (MAE)**
    - Average error between prediction and observation
    - $$MAE = \frac{1}{n} \sum_{i=1}^{n} |e_i|$$
  - **Root Mean Squared Error (RMSE)**

- Used for regression - **measures standard deviation of errors made by a system**
- N - instances in dataset
- **Sensitive to outliers**
-
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} = \sqrt{\frac{e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2}{n}}$$

- **Online**
    - Measures success on live data
    - Often uses metrics like click-through rates, conversion rate, likes etc
    - **Online Evaluations**
        - **A/B Tests**
            - Showing variation A to 50% of users, and B to another 50%
            - **Interleaving**
                - Rankings are mixed, all kind of variations
                - **Random**
                    - Randomly show the user A or B
                - **Top k**
                    - Show the first k A, and the second K B
                - **Fixed amount**
                    - Switch every n user
            - **Pros**
                - Most relevant evaluation
            - **Cons**
                - Time consuming
                - Limited number of tests possible
                - May rely on suboptimal ground-truths and metrics
        - **Splitting Methods**
            - **Hold-out validation**
            - **K-fold cross validation**
                - Split data into k equally sized blocks
                    - Typically k = 10
                - Conduct k evaluations
                    - Train on fold 1..k, evaluate on remaining
                - **Sensible for small datasets**
                    - **Increases run time by factor of k**

            - Better performance than X/Y split
        - **Leave-one-out cross validation (LOOCV)**
        - **Bootstrapping**
        - **Stratified Sampling**
        - **Monte Carlo Cross Validation**
    - **Importance of Time**
        - **Rating performance over time**
            - Typically just report the performance score
            - Assumes that it will perform the same in the future
            - **Not realistic - performance changes hugely over time**
        - **Consider time in training and testing**
            - Very important
            - **Easy to predict current values if we know the future** (e.g. know full route of journey, we will know our current optimal move)
            - E.g. predict bitcoin price looking at samples taken from the last 5 years vs predict bitcoin price looking at price from 2011 to early 2017
                - Won't predict the spike with latter
            - **Time matters**
            - **Not needed in most applications (recommenders, recognition etc)**
- **Live data, and metrics can be used for offline evaluation also**

- **Beyond Offline Evaluation**
    - **(Business) Objectives**
        - Max profit, min costs, win max # users, max user satisfaction, have the best product (most effective, cheapest, value)
        - **Beyond accuracy**
            - Minimize harm
            - Serendipity
            - Diversity
            - Novelty
            - Coverage
        - **KPIs relate directly to business goals**
            - views/referrals/retweets/likes - brand awareness
- **Distribution Shift**
    - **Assumption of offline evaluation**
        - Data is stationary
        - Models that perform well in offline will perform well in online
        - **Not always realistic**

        - **Particularly big problem in research**
            - Working with old data
            - Assume the findings will generalize to environments with different data
        - **Can be measured by comparing the difference between offline evaluation and online evaluation performance**
        - **Large discrepancy -> update offline data, retrain or re-engineer**

- **Baselines**
    - Baselines give the results meaning, allowing for comparison
        - **Can compare globally (in research communities) or locally (within your company/system)**
    - **Without a baseline, performance assessments of algorithms are of little relevance**
    - E.g. my algorithm gets 86% accuracy - is that good or bad?
- **Ground Truth**
    - **"Real truth" can rarely be measured**
    - So we infer/approximate a ground truth
    - **Best measure available**
        - E.g. witnesses in a trial, purchase history to suggest taste/satisfaction
    - **Difficult to find**
    - **Problems with ground truths**
        - General noise, assumes that the examples/problem environment is somewhat perfect (citation recommender assumes authors have cited all relevant papers), can only perform to the standard of the ground truth
- **Gold Standard**
    - Analogous to monetary gold standard that allows comparing currency values
    - **Best method or data (under reasonable conditions)**
        - Data: dataset with the most accurate ground-truth
        - Method: Best performing method
    - **Medical Example (Method)**
        - Ideal test method: autopsy
        - Gold Standard Method: x-ray (*the best alternative method which keeps the patient alive*)
    - **ML example (Data)**
        - Many datasets (ground truths) annotate the relevance of docs and search queries
        - One best dataset (gold standard)
    - **Makes different evaluations roughly comparable**
    - **<u>7 myths about ground truths and gold standards</u>**
        - Most data collection efforts assume there is one correct interpretation for every input example
        - To increase the quality of annotation data, disagreement amongst annotators should be avoided or reduced

            - When specific cases continuously cause disagreement, more instructions are added to limit interpretations
            - Most annotated examples are evaluated by one person
            - Human annotators with domain knowledge provide better annotated data
            - The maths of using ground truths treats every example the same - correctness is boolean
            - Once human annotated data is collected for a task it is used over and over again with no updates

- **Significance & Co.**
    - **Statistical significance**
        - **Describes probability that an observed distance is caused by chance**
        - **Typical p value** should be **smaller** than **0.05 or 0.01**
        - Statistically insignificant results are of little value
        - **Statistically significant results can still be false or insignificant**
        - "**The p-value gives information about the probability of obtaining evidence. It doesn't quantify the strength of the evidence**"
    - **P-Hacking**
        - "If you torture your data long enough it will confess"
        - Can have a statistically significant result for cats landing on all fours, but it's irrelevant
        - Can also have statistically significant results which aren't reproducible