Lecture 2:

Parameters: Things that you learn during trainings

The decision tree algorithm primarily learns two things:

1.the best feature to split on

2.the threshold for the feature to split on at each node

These are called parameters of the decision tree model. When predicting on new examples, we need parameters of the model.

Hyper-parameters: Things that you set to control the training

**Decision trees:**

~ We can also use decision tree for regression problems.

<https://ubc-cs.github.io/cpsc330/lectures/02_decision-trees.html#decision-tree-for-regression-problems>

~ Decision stump: A decision tree with only one stump (5:30). Not necessarily all leaf nodes are pure in this case.

~ If we let the decision tree grow without limit, we does NOT always end up with pure leaf nodes! (Imagine two features have the exactly same value, then how to separate them?)

Lecture 3 ML Fundamentals

Fundamental tradeoff of supervised learning:

~ As you increase the model complexity, E\_train tends to go down but E\_valid - E\_train tends to go up (The gap goes up but not exactly E\_valid goes up!!!!).

Also bias vs variance tradeoff (only either one will be high)

**Bias (high=underfitting):**

**the tendency to learn the same wrong thing or failing to learn something important**

**Variance (high=overfitting):**

**the tendency to learn random things irrespective of the real sign**

Lecture 4 knn

**k-NN:**

~ k-nn selects no feature at all, it's gonna use up all features to calculate the euclidean tree (The decision tree on the other hand may be working at some small sets of a feature) (E.g. a tree stump (with depth = 1) has one split and 1 good feature)

~ training error is 0 when k = 1 (always find the closest one which is itself!)

~ We can also use knn to do regression problem (take the average of k neighbours)

Pros of k-NNs for supervised learning¶

~ Easy to understand, interpret.

~ Simple hyperparameter k (n\_neighbors) controlling the fundamental tradeoff.

~ Can learn very complex functions given enough data.

~ Lazy learning: Takes no time to fit

Cons of k-NNs for supervised learning¶

~ Can be potentially be VERY slow during prediction time, especially when the training set is very large.

~ Often not that great test accuracy compared to the modern approaches.

~ It does not work well on datasets with many features or where most feature values are 0 most of the time (sparse datasets).

Curse of dimensionality (More features might result in bad results for k-nn!):

~ Affects all learners but especially bad for nearest-neighbour.

~ k-NN usually works well when the number of dimensions d is small but things fall apart quickly as d goes up.

~ If there are many irrelevant attributes, k-NN is hopelessly confused because all of them contribute to finding similarity between examples.

~ With enough irrelevant attributes the accidental similarity swamps out meaningful similarity and k-NN is no better than random guessing.

**SVM RBF (Support Vector Machine with RBF kernel)**

Superficially, SVM RBFs are more like weighted k-NNs.

~ The decision boundary is defined by a set of positive and negative examples and their weights together with their similarity measure.

~ A test example is labeled positive if on average it looks more like positive examples than the negative examples.

The primary difference between k-NNs and SVM RBFs is that

~ Unlike k-NNs, SVM RBFs only remember the key examples (support vectors). So it’s more efficient than k-NN.

~ SVMs use a different similarity metric which is called a “kernel” in SVM land. A popular kernel is Radial Basis Functions (RBFs) They usually perform better than k-NNs!

Support vectors¶

~ Each training example either is or isn’t a “support vector”.

~ This gets decided during fit.

~ Main insight: the decision boundary only depends on the support vectors.

Key hyperparameters of rbf SVM are

~ gamma (larger gamma → more complex)

~ C (larger C → more complex)

Lecture 5 – Preprocessing

1. knn will generally suffer from problem like: the eculidean distance might be completely dominated by the features with super large values. Features on a smaller scale will be ignored and can be highly informative, there is no reason to ignore them. (Though scaling could alleviate this problem
2. That's why we need to scale/transform the data
   1. StandardScaler() (a.k.a "standardisation") in scikit-learn: sets sample mean to 0, s.d. to 1. e.g. X -= np.mean(X, axis=0), X /= np.std(X, axis=0)
   2. MinMaxScaler() (a.k.a "normalisation") in scikit-learn, sets range to [0,1], e.g. X -= np.min(X, axis=0), X /= np.max(X, axis=0)

Lecture 7 – Linear models

~ Ridge is similar to linear regression but with a hyperparameter called alpha. When alpha = 0, they are the same.

larger alpha → likely to underfit

smaller alpha → likely to overfit

Possible alpha range: [0.001, 0.01, 0.1, 0, 10 100]

Logistic Regression has similar hyperparams with SVC

smaller C → might lead to underfitting

bigger C → might lead to overfitting

Possible C range: [0.001, 0.01, 0.1, 0, 10 100]

~pass "handle\_unknown="ignore"" argument to OneHotEncoder

It creates a row with all zeros if we hit a situation that the test set has a feature that is not in the train set

Lecture 8 – Hyperparameter optimization

Advantages of RandomizedSearchCV¶

~ Faster compared to GridSearchCV.

~ Adding parameters that do not influence the performance does not affect efficiency.

~ Works better when some parameters are more important than others.

Overfitting of the validation error¶

Why do we need to evaluate the model on the test set in the end?

Why not just use cross-validation on the whole dataset?

While carrying out hyperparameter optimization, we usually try over many possibilities.

If our dataset is small and if your validation set is hit too many times, we suffer from optimization bias or overfitting the validation set.

Lecture 9 – classification metrics

1.Precision: Among the positive examples you identified, how many were actually positive? precision = TP / (TP + FP)

2.Recall: Among all positive examples, how many did you identify?, which is equal to TP / (TP + FN) = TP / #positives

3.F1 score: F1-score combines precision and recall to give one score, which could be used in hyperparam optimization.

f1 = 2 \* (precision \* recall) / (precision + recall)

* One way to do this is by computing the area under the PR curve.
* This is called **average precision** (AP score)
* AP score has a value between 0 (worst) and 1 (best).

AP vs. F1-score¶

It is very important to note this distinction:

F1 score is for a given threshold and measures the quality of predict.

AP score is a summary across thresholds and measures the quality of predict\_proba.

1.PR(Precision-recall) curve (x-axis is precision and y-axis is recall): AP score is the area under the PR(Precision-recall) curve

~ AP score (Average precision) is independent of the threshold because it actually goes through every single threshold to plot the curve but other metrics are subject to threshold changing (E.g. Recall)

2.receiver operating characteristic (ROC) curve: more popular (FPR-x vs TPR-y) (Top left best)

it considers all possible thresholds for a given classifier given by predict\_proba but instead of precision and recall it plots false positive rate (FPR) and true positive rate (TPR or recall).

3.AUC (Area under the curve): evaluating the ranking of positive examples.

1.0: all positive points have a higher score than all negative points

0.5: means random chance

AUC can be interpreted as evaluating the ranking of positive examples.

What’s the probability that a randomly picked positive point has a higher score according to the classifier than a randomly picked point from the negative class.

Lecture 10 – regression metrics

Alpha in Ridge(): higher = simpler model, lower = more complex model, which is exactly the reverse of C in linearRegression

4. MAPE: percent\_errors = (pred\_train - y\_train) / y\_train \* 100.0