ECE M146 Introduction to Machine Learning

Prof. Lara Dolecek

ECE Department, UCLA

Today: Review

- Today, we are going to pause and summarize the concept we learnt thus far.

Types of learning

- Supervised learning

- Unsupervised learning

- Also: semi-supervised learning, reinforcement learning

Types of learning

- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
  - Clustering
  - Dimensionality reduction
  - Density estimation

- Also: semi-supervised learning, reinforcement learning

Types of learning

- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
  - Clustering
  - Dimensionality reduction
  - Density estimation

- Also: semi-supervised learning, reinforcement learning

Approaches to supervised learning

1. Use a discriminant function:

2. Discriminative modeling:

3. Generative modeling:

What is classification ?

- Binary classification:

- Multiclass classification:

- Using a binary classifier to construct a multiclass classifier:
  - One vs. All
  - All vs. All

What is regression ?

Methods we have learnt thus far

1. Perceptron
2. Linear least squares for linear regression
3. Logistic regression
    - Binary: logistic sigmoid
    - Multiclass: softmax
4. Decision Trees
5. K-NN
6. SVM

Parametric methods

- Which ones from the previous list are parametric methods ?

- The typical objective is to find a vector w of the same dimension as a data point x; output is governed by a function of $w^T x$.
- The dimension of w (and the complexity of the model) do not depend on the number of data points N.
- Essentially, vector w (or vector w and scalar b, if latter is modeled explicitly) tells you everything you need to know about the model.

Non-parametric methods

- Which ones from the previous list are non-parametric methods ?

- There is no mathematical formula describing the decision boundary.
- Complexity of the model grows with the number of the data points.

Use of a discriminant function

- Which ones from the previous list use a discriminant function -- a function that maps an input data point directly to the output (e.g., to the class label in classification)?

Discriminative models

- Which ones from the previous list are discriminative models ?

Linear models

- Which ones from the previous list are linear models ?

Regularization techniques

- Regularization – penalty term added to the loss function to suppress overfitting.
- L1 loss:
- L2 loss:

Model assessment techniques

- Use validation:

- Use cross validation:

Kernel techniques

- Can operate in higher dimensional space – where the data is linearly separable – without incurring dimensionality complexity penalty.
- Replace inner product by the inner product of the feature maps, without computing them explicitly.

Kernel techniques in action

Mathematical tools

Concept 1: Minimization or maximization of a function.

- Function is loss or (log) likelihood.

- Take a derivative and set it to zero.
- If not possible, do gradient descent or ascent.

Mathematical tools

Concept 2 : Matrix calculus:

Mathematical tools

Concept 3: Optimization

- Formulate a constrained optimization problem
- Convert a constrained optimization problem into an unconstrained problem via Lagrangian that incorporates constraints.
- From primal to dual.
- Conditions for convex problems.

Perceptron

- On-line algorithm for binary classification of linearly separable data.
- Update equation:

- Can also be interpreted as stochastic gradient descent.

- Has a mathematical proof of convergence.

Linear regression

- Goal is to minimize loss function:
- Matrix format:
- Set derivatives to zero (matrix calculus):
- Can also do as gradient descent:

Logistic regression

- Method for binary classification. We switched from y being in {+1,-1} to y in {0,1} for mathematical convenience.
- Discriminative modeling:
- Maximize likelihood:
- Gradient descent:
- Multi-class classification via softmax and max. likelihood

Decision Trees

- Non-parametric but interpretable method; can be used for classification or regression.
- Tree is built recursively:
- For binary classification, a principled way to split is via information gain:
- Stopping criteria.
- How to prevent overfitting.

KNN

- Non-parametric as well, but simple and intuitive; can be used for classification or regression
- Use K odd. K=1 is the simplest yet does reasonably well.
- Nothing really to do at training time. Procedure at test time:
  - For K =1, compute all squared distances and find the training point closest to the test point and assign its label to be the label of the test point.
  - For K =3, compute all squared distances and find the 3 training points closest to the test point and assign the label owned by the majority to be the label of the test point.
- Different types of distances.
- Choice of K.

SVM

- Explicitly model the margin.

- After some mathematical derivations, arrive at the constrained optimization problem:

- Convert to unconstrained optimization.
- Use primal to dual transformation. Optimize by taking derivatives.
- Support vectors.

Soft SVM

- Allow for slack.
- This introduces additional parameters, which in turn enlarges the optimization problem:

- Increased number of support vectors.

How do they compare ?

- Perceptron vs. SVM

- Soft SVM vs. logistic regression

- Linear least squares for classification vs. logistic regression/ soft SVM