ECE M146 Introduction to Machine Learning

Prof. Lara Dolecek

ECE Department, UCLA

Today: Review

- Today, we are going to review the concepts covered in this course, with the emphasis on the second half.

Methods we learned in this course

- Perceptron
- Linear least squares for linear regression
- Logistic regression
  - Binary: logistic sigmoid
  - Multiclass: softmax
- Decision Trees
- K-NN
- SVM

Methods we learned in this course

- Perceptron
- Linear least squares for linear regression
- Logistic regression
  - Binary: logistic sigmoid
  - Multiclass: softmax
- Decision Trees
- K-NN
- SVM

- Naïve Bayes Classifier
- Gaussian Discriminant Analysis
  - Linear and quadratic DA
- K-means
- PCA
- Expectation-Maximization
- Ensemble methods
  - Bagging; Boosting; AdaBoost

Supervised Learning

- Training data is labeled.

- At test time, decide class membership in classification (often binary, but can be multi-class) or assign real-valued label in regression.

- Which of the previous methods perform supervised learning ?

Methods we learned in this course

- Perceptron
- Linear least squares for linear regression
- Logistic regression
  - Binary: logistic sigmoid
  - Multiclass: softmax
- Decision Trees
- K-NN
- SVM
- Naïve Bayes Classifier
- Gaussian Discriminant Analysis
  - Linear and quadratic DA
- K-means
- PCA
- Expectation-Maximization
- Ensemble methods
  - Bagging; Boosting; AdaBoost

Unsupervised Learning

- Training data is not labeled.

- At test time, goal it so organize the data (partition into groups/clusters, to project, to compress).

- Which of the previous methods perform unsupervised learning ?

Methods we learned in this course

- Perceptron
- Linear least squares for linear regression
- Logistic regression
  - Binary: logistic sigmoid
  - Multiclass: softmax
- Decision Trees
- K-NN
- SVM
- Naïve Bayes Classifier
- Gaussian Discriminant Analysis
  - Linear and quadratic DA
- K-means
- PCA
- Expectation-Maximization
- Ensemble methods
  - Bagging; Boosting; AdaBoost

Back to classification

- Discriminant function

- Discriminative models

- Generative models

Generative models

- Which ones from the previous list are generative models ?

Naïve Bayes Classifier

- The key assumption is that the features are independent given the class label.
- We considered the following set up, with features being binary (but this framework can of course be applied to other distributions including multinomial, Gaussian etc.).
- Maximize:

Naïve Bayes Classifier

Gaussian Discriminant Analysis

- Individual classes are modeled as Gaussians. The objective is to maximize the overall likelihood.

GDA as Naïve Bayes

- When GDA model additionally satisfies the conditions of the Naïve Bayes model i.e., the features are conditionally independent given the class label, the following holds for the Gaussians:

Unsupervised learning

Clustering

- K-means algorithm
- Expectation maximization for clustering

Dimensionality reduction

- PCA

K-means clustering

- We are given unlabeled data that that we want to organize into clusters, such that each cluster prototype is the best representative of the data points within its cluster.

- Mathematically, we seek to minimize the following distortion measure:

K-means clustering

- Minimization of the preceding expression is done through an iterative optimization process, as follows:

- Initialize prototypes.

- Then, iterate between the two steps:

1. Find the best indicators (argmin function) for fixed prototypes.
2. Refit prototypes for fixed indicators.

- Terminate when the target cost function is reached.

Expectation Maximization (EM)

- Cluster modeling:

- Data likelihood:

- The goal as usual is to maximize this data (log) likelihood. Cannot take the derivatives directly, so we iterate.

Expectation Maximization (EM)

- Maximization of the preceding expression is done through an iterative optimization process, as follows:

- Initialize model parameters.
- Then, iterate between the two steps:

1. Evaluate responsibilities (posterior probabilities) given the current model values.
2. Perform MLE estimate on model parameters (refitting) given fixed responsibilities.

- Terminate when the target function is reached.

EM and K-means

- Recall that K-means algorithm provides “hard” cluster assignments (0 or 1).
- EM with Gaussian Mixture Model provides “soft” cluster assignment where a point can have fractional membership across multiple clusters.

- EM is a general technique for finding maximum likelihood solution with hidden variables.

PCA

- For a given data set of dimension D, we seek to project into a lower dimensional space of dimension M, M < D.

- The projection that is most informative is the direction (subspace) with the highest variance of the projected data.

- For D= 1, set up this problem as a constrained optimization problem and use Lagrangian formulation to solve. Arrive at the eigen vector of the largest eigen value.

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.

- Used where ?

Mathematical tools

Concept 2 : Matrix calculus

- Derivatives of a scalar with respect to the vector, matrix that extend the “usual” definition of derivatives.

Used where ?

Mathematical tools

Concept 3: Constrained 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.

- Used where ?

Mathematical tools

- Concept 4: Probability
- Compute total probability, conditional probability, marginal probability.
- Characterize and manipulate common distributions, including Bernoulli and Gaussian.
- Derive Mixture Distribution e.g., GMM

- Used where ?

Mathematical tools

Concept 5: Linear algebra

- Projections, vectors norms
- Eigen vectors and eigen values
- Positive (semi) definite matrices

- Used where ?

Combining weak learners

- Bagging (bootstrap aggregation) – is a parallel approach where we generate new samples by sampling with replacement and train classifiers in parallel on these data sets.

- Boosting – serial approach where we reweigh difficult data points (misclassified points) at the input for the next classifier.

- AdaBoost – is a popular method that performs boosting in a principled way. We showed that it minimizes exponential loss.

General techniques

- Regularization – add a penalty term to avoid overfitting. Common choices are L1 or L2 norm.

- Kernel techniques – allow us to operate in a high dimensional space at the complexity level of a low dimensional space by replacing inner products by the inner products of their maps, without evaluating these maps explicitly.

- Model assessment and validation – procedures for assessing how well will our model perform on unseen data; procedures for selecting the best choice of a hyperparameter.

- Thank you!