CAP 5610: Machine Learning

Lecture:
Midterm Review

Instructor: Dr. Gita Sukthankar
Email: gitars@eecs.ucf.edu

Schedule

- Midterm exam next Tuesday Oct 22 (25% of final grade)
  - You are allowed to bring one page of notes covered front and back (cannot be shared)
  - No programming questions
  - No reading questions
  - No questions about research applications other than Eigenfaces

Example Questions

True or false: Assuming that the training set is the same, the cluster means found by k-means clustering will be the same across multiple repetitions of the algorithm.

Given the conditional probabilities of a set of a features for a naive Bayes classifier, decide on a class label.

Naïve Bayes Example

Probability	positive	negative
P(*Y*)	0.5	0.5
P(small | *Y*)	0.4	0.4
P(medium | *Y*)	0.1	0.2
P(large | *Y*)	0.5	0.4
P(red | *Y*)	0.9	0.3
P(blue | *Y*)	0.05	0.3
P(green | *Y*)	0.05	0.4
P(square | *Y*)	0.05	0.4
P(triangle | *Y*)	0.05	0.3
P(circle | *Y*)	0.9	0.3

We learn these probabilities from the training data.

Test Instance:
<medium ,red, circle>

Naïve Bayes Example

Probability	positive	negative
P(*Y*)	0.5	0.5
P(medium | *Y*)	0.1	0.2
P(red | *Y*)	0.9	0.3
P(circle | *Y*)	0.9	0.3

Test Instance:
<medium ,red, circle>

Answer:
Drawn from the positive urn

P(positive | *X*) = P(positive)*P(medium | positive)*P(red | positive)*P(circle | positive) / P(*X*)
0.5 * 0.1 * 0.9 * 0.9
= 0.0405 / P(*X*) = 0.0405 / 0.0495 = 0.8181

P(negative | *X*) = P(negative)*P(medium | negative)*P(red | negative)*P(circle | negative) / P(*X*)
0.5 * 0.2 * 0.3 * 0.3
= 0.009 / P(*X*) = 0.009 / 0.0495 = 0.1818

P(positive | *X*) + P(negative | *X*) = 0.0405 / P(*X*) + 0.009 / P(*X*) = 1

P(*X*) = (0.0405 + 0.009) = 0.0495

For purposes of making a decision, we can ignore the denominator since it is the same for both classes.

Lecture 1: Introduction

- Feature vector
- K nearest neighbor
- Distance functions: Euclidean, Manhattan
- Training/test set split
- K-fold cross validation and Leave one out
- Problems of overfitting
- Curse of dimensionality definition
  - Do not need to know proof
- Do not need to know: definitions of ML, oracle function, applications

Lecture 2: Probability Theory

- Do not need to know history of ML, beta distribution as conjugate prior, KPark crowdsourcing example
- Probability: discrete and continuous
- Distributions: binomial, Bernoulli, Gaussian
- Sample mean and variance
- Conditional probability and Bayes rule
- Joint probability tables
- Conditional independence
- MLE estimate for Bernoulli and univariate Gaussian
- MAP: pseudocounts and how it relates to MAP

Lecture 3 and 4: Bayesian Classifiers

- Prior, class-conditional, and posterior probability distributions
- Maximum a posteriori decision rule and likelihood ratio
- Definition of Bayes error
- Naive Bayes classifier example
- Gaussian class conditional density (1D) and special case (equal covariance matrix)
- Bayes classifiers for continuous feature
  - MLE
  - MAP
- Evaluation: true/false positive/negative, precision, recall, confusion matrix
- Do not need to know: nearest neighbor error, Dirichlet distribution, graphical model

Lecture 5: Logistic/Linear Regression

- Difference between discriminative and generative models
- Logistic function for P (Y=1|X) $= \frac{1}{1 + \exp(\sum_{i=1}^{N} \omega_i X_i + \omega_0)}$
- Multiclass model

$$P(Y = c|X) = \frac{\exp(\sum_{i=1}^{N} \omega_{ci} X_i + \omega_{c0})}{\sum_{c'=1}^{C} \exp(\sum_{i=1}^{N} \omega_{c'i} X_i + \omega_{c'0})}$$

- Role of regularizer to avoid overfitting
- Comparison between Naive Bayes and logistic regression (parameter reduction)
- Idea behind bias-variance decomposition
- Difference between linear and logistic regression
- MLE pseudo inverse solution for linear regression (not the MAP)

Lecture 6 and 7: Support Vector Machines

- Meaning of terms support vector and margin
- Margin example for $wx+b>1$ and $wx+b<-1$
- Formulation of optimization problem
- Hard margin vs. soft: use of slack variables
- Kernel trick
- Different kernel types and choosing kernels
- SVM multiclass
- Weighted SVMs
- Don’t need to know optimization details, KKT conditions, example applications, regression, SMO, Platt scaling

Lecture 8 and 9: Neural Networks and CNNs

- Neural network structure
- Stochastic gradient descent
- Backpropagation
- Momentum
- Purpose of regularized square error (but not the formulas)
- NN vs. SVM comparison
- Autoencoder structure
- Advantages of deep learning: feature detection in neural networks
- Convolution and max pooling operations
- Do not need to know self-driving applications, details of LeNet5, or the XOR example

Lecture 10: Dimensionality Reduction

- Purpose of dimensionality reduction
- PCA
- Eigenface example
- FDA
- Do not need to know: kernel PCA

Lecture 11: Decision Trees

- Procedure for learning decision trees
- Attribute choice using conditional entropy
- Techniques for avoiding overfitting decision trees
- General idea of random forest but not specifics of algorithm

Lecture 12: Boosting

- Bootstrap estimation and bagging
- Adaboost (algorithm)
- Do not need to know the training error reduction proof
- Do not need to know the Viola-Jones face detector

Lecture 13: Clustering

- K-means clustering algorithm
- Problems with k-means
- Calculating cluster purity
- Do not need to know convergence proof, Rand index, research applications (SocialSim),