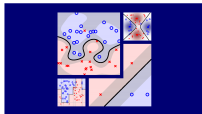


Machine Learning Techniques (機器學習技法)



Lecture 1126: Summary

Hsuan-Tien Lin (林軒田)

`htlin@csie.ntu.edu.tw`

Department of Computer Science
& Information Engineering

National Taiwan University
(國立台灣大學資訊工程系)



Agenda

Lecture 1126: Summary

Model 1: PLA

- err: 0/1
- $\widehat{\text{err}}$: $\max(-ys, 0)$
- optimization:
 - SGD on $\widehat{\text{err}}$
 - special minimizer of err when linear separable
- Φ : identity
- regularization/validation: none
- parameters: none
- practical use: online learning, and ‘teaching’

Model 2: pocket

- err: 0/1
- $\widehat{\text{err}}$: none
- optimization: special random search with PLA as core
- Φ : identity
- regularization/validation: none
- parameters: number of iterations T
- practical use: seldom

Model 3: ridge linear regression

- err: squared
- $\widehat{\text{err}}$: squared
- optimization: analytic solution
- Φ : identity
- regularization/validation: L2 regularization
- parameters: regularization level λ
- practical use: for 'decent' baseline solution

Model 4: logistic regression

- err : cross entropy
- $\widehat{\text{err}}$: cross entropy
- optimization: GD/SGD (basic)
- Φ : identity
- regularization/validation: often L2 regularization, or L1 if needing sparsity
- parameters: regularization level λ and GD/SGD iterations T , step size η
- practical use: very useful for baseline hard/soft classification

Model 5: ridge polynomial regression

- err: squared
- $\widehat{\text{err}}$: squared
- optimization: analytic solution
- Φ : polynomial
- regularization/validation: L2 regularization
- parameters: regularization level λ and polynomial degree Q
- practical use: often for (1-D) regression problems

Model 6: soft-margin linear SVM

- err: 0/1
- $\widehat{\text{err}}$: hinge
- optimization: (special) quadratic programming
- Φ : identity
- regularization/validation: large-margin/leave-one-out bound or CV
- parameters: error penalty rate C
- practical use: very popular for large-scale classification

Model 7: soft-margin kernel SVM

- err: 0/1
- $\widehat{\text{err}}$: hinge
- optimization: (special) quadratic programming
- Φ : embedded in kernel K
- regularization/validation: large-margin/leave-one-out bound or CV
- parameters: error penalty rate C and kernel parameters
- practical use: very popular for mid-sized classification

Model 8: kernel logistic regression

- err : cross entropy
- $\widehat{\text{err}}$: cross entropy
- optimization: GD/SGD
- Φ : embedded in kernel K
- regularization/validation: L2 regularization
- parameters: regularization level λ and kernel parameters
- practical use: Probabilistic SVM more often used

Model 9: kernel ridge regression

- err: squared
- $\widehat{\text{err}}$: squared
- optimization: analytic solution (but often solved iteratively)
- Φ : embedded in kernel K
- regularization/validation: L2 regularization
- parameters: regularization level λ and kernel parameters
- practical use: only on very small data

Model 10: AdaBoost

- err: 0/1
- $\widehat{\text{err}}$: exponential
- optimization: functional gradient descent with the help of base algorithm
- Φ : diverse hypotheses found iteratively
- regularization/validation: often through early stopping
- parameters: number of iterations T
- practical use: 'boost' decision trees/stumps

Model 11: Decision Tree

- err: 0/1 or squared
- $\widehat{\text{err}}$: not fully clear
- optimization: heuristic greedy
- Φ : conditional hypotheses found recursively
- regularization/validation: pruning
- parameters: lots of heuristic choices
- practical use: 'explainable' nonlinear model

Model 12: Bagging/Random Forest

- err : squared, or 'any'?!
 - $\widehat{\text{err}}$: squared
- optimization: through base algorithm/decision trees
- Φ : diverse hypotheses through bootstrapping, and random projection/combination
- regularization/validation: variance decreasing/OOB error
- parameters: number of iterations T
- practical use: 'stabilize' any model/tree

Model 13: Gradient Boosted Decision Tree

- err : squared, or any
- $\widehat{\text{err}}$: squared, or any
- optimization: functional gradient descent with the help of base algorithm
- Φ : diverse hypotheses found iteratively
- regularization/validation: often through early stopping
- parameters: number of iterations T
- practical use: very popular for information retrieval and competitions

Model 14: Neural Networks/Deep Learning

- err : squared, or cross entropy
- $\widehat{\text{err}}$: squared, or cross entropy
- optimization: ADAM GD/SGD with help of backprop, Xavier/He initialization, batch normalization (?)
- Φ : learned and represented by hidden neurons
- regularization/validation: dropout, early stopping, sparse connections & weight sharing (e.g. CNN)
- parameters: number of iterations T , step size η , network architecture
- practical use: very popular nowadays for vision/speech

Model 15: RBF Networks

- err: squared, or 0/1
- $\widehat{\text{err}}$: squared
- optimization: prototype-finding with k -means-like algorithms, then analytic solution
- Φ : learned and represented by prototypes
- regularization/validation: number of prototypes
- parameters: number of prototypes k , RBF width γ
- practical use: less popular nowadays

Model 16: Matrix Factorization

- err: squared
- $\widehat{\text{err}}$: squared
- optimization: alternating least squares, or SGD
- Φ : learned hidden factors
- regularization/validation: often just L2
- parameters: number of hidden factors
- practical use: baseline for recommender systems

Summary

Lecture 1126: Summary