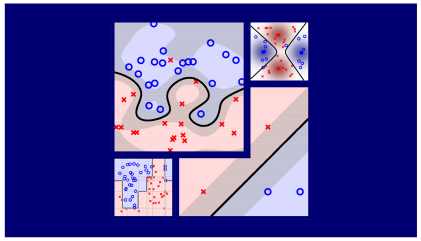
Machine Learning Foundations
(機器學習基石)



Lecture 1126: Super Short 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)

Model 1: PLA

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

Model 2: ridge linear regression

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

Model 3: logistic regression

- err: cross entropy
- $\widehat{\mathrm{err}}$: cross entropy
- optimization: GD/SGD (basic)
- $\mathbf{\Phi}$: 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 4: ridge polynomial regression

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

Model 5: soft-margin linear SVM

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

Model 6: soft-margin kernel SVM

- err: 0/1
- $\widehat{\mathrm{err}}$: hinge
- optimization: (special) quadratic programming
- $\mathbf{\Phi}$: embedded in kernel K
- regularization/validation: large-margin
- parameters: error penalty rate C and kernel parameters
- practical use: very popular for mid-sized classification

Model 7: AdaBoost

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

Model 8: Decision Tree

- err: 0/1 or squared
- $\widehat{\mathrm{err}}$: same as err often
- optimization: heuristic greedy
- $\mathbf{\Phi}$: conditional hypotheses found recursively
- regularization/validation: pruning
- parameters: lots of heuristic choices
- practical use: 'explainable' nonlinear model

Model 9: Bagging/Random Forest

- err: squared, or 'any'?!
- $\widehat{\mathrm{err}}$: squared
- optimization: through base algorithm/decision trees
- $\mathbf{\Phi}$: 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 10: Gradient Boosted Decision Tree

- err: squared, or any
- $\widehat{\mathrm{err}}$: squared, or any
- optimization: functional gradient descent with the help of base algorithm
- $\mathbf{\Phi}$: 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 11: Neural Networks/Deep Learning

- err: squared, or cross entropy
- $\widehat{\mathrm{err}}$: squared, or cross entropy
- optimization: ADAM GD/SGD with help of backprop, Xavier/He initialization
- $\mathbf{\Phi}$: learned and represented by hidden neurons (ReLU / tanh)
- regularization/validation: early stopping, L1/L2, and dropout
- parameters: number of iterations T, step size η, network architecture
- practical use: very popular nowadays for vision/speech