Chapter 3
Supervised Machine Learning (in a nutshell)

supplementary slides to

Machine Learning Fundamentals

©**Hui Jiang 2020**

published by Cambridge University Press

August 2020

CAMBRIDGE
UNIVERSITY PRESS

Machine Learning Pipeline

- data collection
  - the more in-domain data, the better
  - data cleaning
- feature extraction
  - feature engineering
  - feature normalization, dimensionality reduction
- model learning
  - supervised machine learning
  - discriminative vs. generative models

Machine Learning Procedure

1. feature extraction (optional)
2. choose a model from LIST-A
   - discriminative vs. generative models
   - simple vs. complex models
   - parametric vs. non-parametric models
3. choose a learning criterion from LIST-B
   - construct an objective function of model parameters
4. choose an optimization algorithm from LIST-C
   - analytic vs. numerical methods
5. empirical evaluation and (optional) theoretical guarantees
   - whether the learning process converges?
   - how well the learned models generalize?

LIST-A: Machine Learning Models (I)

(i) discriminative models:

- linear models
- bilinear models, quadratic models
- logistic sigmoid, softmax, probit
- nonlinear kernels
- decision trees
- neural networks:
  - FCNN, CNN, RNN, LSTM, transformers, etc.

LIST-A: Machine Learning Models (II)

(ii) generative models:

- Gaussian models
- multinomial models
- Markov chain models
- mixture models
  - Gaussian mixture models
  - hidden Markov models
- entangled models
- deep generative models
  - variational autoencoders
  - generative adversarial nets
- graphical models:
  - Bayesian networks, e.g. naïve Bayes classifiers, LDA
  - Markov random fields, e.g. CRF, RBM
- Gaussian processes

LIST-B: Machine Learning Criteria

- **for discriminative models**:
  - least square error
  - minimum classification error
  - minimum cross-entropy
  - maximum margin
  - minimum L_p norm
- **for generative models:**
  - maximum likelihood
  - maximum conditional likelihood
  - maximum a posteriori
  - maximum marginal likelihood
  - minimum KL divergence

LIST-C: Optimization Methods

- grid search
- gradient descent
- stochastic gradient descent (SGD)
- coordinate descent
- subgradient methods
- Newton's method
- quasi-Newton methods:
  - quickprop, R-prop, BFGS, L-BFGS
- expectation-maximization (EM)
- sequential line search
- alternating direction method of multipliers (ADMM)
- gradient boosting

Case Studies (I)

- not all combinations from the three lists make sense
- **linear regression**
  (linear model) $\times$ (least square error) $\times$ (closed-form or gradient descent)
- **ridge regression**
  (linear model) $\times$ (least square error $+$ min L_2 norm) $\times$ (closed-form or gradient descent)
- **LASSO**
  (linear model) $\times$ (least square error $+$ min L_1 norm) $\times$ (subgradient descent)

Case Studies (II)

- **logistic regression**
  (linear model + logistic sigmoid) $\times$ (max likelihood) $\times$ (gradient descent)
- **linear SVM**
  (linear model) $\times$ (max margin) $\times$ (gradient descent)
- **nonlinear SVM**
  (nonlinear kernels + linear model) $\times$ (max margin) $\times$ (gradient descent)
- **soft SVM**
  (kernels + linear model) $\times$ (min linear error + max margin) $\times$ (gradient descent)

Case Studies (III)

- **matrix factorization**
  (bilinear model) $\times$ (least square error + min L_2 norm) $\times$ (gradient descent)
- **dictionary learning**
  (bilinear model) $\times$ (least square error + min L_1 norm) $\times$ (gradient descent)
- **topic modelling**
  (latent Dirichlet allocation) $\times$ (max marginal likelihood) $\times$ (EM algorithm)
- **deep learning**
  (neural networks) $\times$ (min cross-entropy error) $\times$ (SGD)
- **boosted trees**
  (decision trees) $\times$ (least square error) $\times$ (gradient boosting)