Intelligent Systems

1 Introduction 10/03

- Three types of machine learning
  - supervised learning
  - unsupervised learning
  - reinforcement learning

2 Unconstrained optimization problem 10/10

- convex optimization problem
  - $\boldsymbol{x}^* = argmin_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$
  - $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ is the necessary condition
- Steepest descent method
  - use $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \varepsilon_k \nabla f(\boldsymbol{x}_k)$ and find the suitable x
  - In order to find the step distance ϵ, two ways are used
    - exact line search, which is to find $\min_{\varepsilon_k > 0} f(\boldsymbol{x}_k - \varepsilon_k \nabla f(\boldsymbol{x}_k))$
    - backtracking line search, which is to decay ϵ_k until it satisfies almiho rule,

$$f(\boldsymbol{x}_k - \varepsilon_k \nabla f(\boldsymbol{x}_k)) - f(\boldsymbol{x}_k) \leq -\alpha \varepsilon_k \|\nabla f(\boldsymbol{x}_k)\|^2$$

- Newton method
  - use $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \varepsilon_k \left(\nabla^2 f(\boldsymbol{x}_k)\right)^{-1} \nabla f(\boldsymbol{x}_k)$
  - However, this algorithm takes a lot of time to calculate the invertible matrix.
- Quasi-Newton method
  - It's practical because it does not include the calculation of the invertible matrix.

3 Constrained optimization problems 10/17

3.1 $\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$ subject to $\boldsymbol{g}(\boldsymbol{x}) = \mathbf{0}$

- penalty method
- method of Lagrange multipliers
  - let L be $L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x})$ and solve the equation $\nabla_\lambda L(x, \lambda) = 0$ and $\nabla_x L(x, \lambda) = 0$

- dual ascent method
  - method of lagrange multipliers needs complex calculations, so separate it into two steps
  - First, calculate $\boldsymbol{x}_{k+1} = argmin_{\boldsymbol{x}\in\mathcal{X}} L\left(\boldsymbol{x}, \boldsymbol{\lambda}_k\right)$ then do $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \varepsilon_k \boldsymbol{g}\left(\boldsymbol{x}_{k+1}\right)$
- method of multipliers
  - use augmented Lagrangian instead of Lagrangian
  - augmented Lagrangian is defined as $L_c(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}) + \frac{c}{2}\|\boldsymbol{g}(\boldsymbol{x})\|^2$

3.2 $\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ subject to $\boldsymbol{h}(\boldsymbol{x}) \leq \boldsymbol{0}$

- penalty method
- KKT conditions
  - Seeing an active constraint and an inactive constraint, we gain KKT conditions.
- If we take dual problems, which often be called Lagrange dual problems, we could reduce the number of constraints.

4 Searching 10/24

- Translate some conditions into state-space
- Blind search (no cost)
  - DFS and BFS
  - iterative deeping search
- cost search
  - greedy search
    - relation with DFS
  - optimal search (Dijkstra's algorithms)
    - relation with BFS
  - Two above methods are blind search, but if you know something in the graph, you can use heuristics search like A-star search
- game tree
  - min-max
  - alpha-beta

5 Probability distribution 10/31

- cumulative distribution function
- skewness and kurtosis
- moment generating function
- convolution

6 Various kinds of probability distribution 11/14

6.1 Discrete probability distribution

- discrete uniform distribution
- binomial distribution
  - the number of successes in a sequence of n independent experiments, which says yes with probability p
- hypergeometric distribution
  - the probability of k successes in n draws without replacement
- Poisson distribution
  - the probability of a given number of events occurring in a fixed interval of time or space

6.2 Continuous probability distribution

- gaussian distribution
  - $f(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{(s-\mu)^2}{2\sigma^2})$
- gamma distribution
  - The time which happens the first event under the probability of a given number of events occurring in a fixed interval of time or space
- beta distribution

7 Law of large numbers and central limit theorem 11/21

7.1 some inequalities

- Chebyshev's inequality
- Markov's inequality, Jensen's inequality, and Holder's inequality

7.2 law of large numbers

- weak law of large numbers
- strong law of large numbers

7.3 Central limit theorem

- if n is large, $\bar{X_n}$ is folloed gaussian distribution

7.4 hypothesis testing

8 supervised learning 11/28

8.1 regression

- least squares $\min_{\boldsymbol{\theta}} \left[\frac{1}{2}\sum_{i=1}^{n}\left(y_i - f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)\right)^2\right]$
- using kernel matrix, $\frac{1}{2}(\boldsymbol{K\theta} - \boldsymbol{y})^\top(\boldsymbol{K\theta} - \boldsymbol{y})$
- To prevent overfitting, normalization are used
- $\frac{1}{2}\sum_{i=1}^{n}\left(y_i - f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{n}\theta_j^2$
- cross-validation

8.2 classification

- margin and svm
- if the data cannot be diveded with the hyperplane, we use $\boldsymbol{\phi(x)}$ to project feature space, which can be diveded with hyperplane.
- kernel trick, which use $\boldsymbol{\phi(x)}^\top\boldsymbol{\phi(x)}$ instead of $\boldsymbol{\phi(x)}$
- subgradient method is used to this optimization

9 unsupervised learning 12/05

- principal component analysis
  - $T_{best} = argmintr(TCT^\top)$
  - relation with eigen value problems
- kmeans clastering
  - well known algorithms for clastering
  - kernel trick
- some applications of PCA and kmeans

10 sequential data 12/12

- introuction to NLP
  - structured prediction
  - sequence labeling
- Hidden Markov Model
  - Viterbi algorithms and dynamic programming

11 HMM 12/19

- parameter inference
  - maximum likelihood analysis
  - Baum-Welch algorithm

12 Parsing 12/26

- parse tree
- CFG
- CKY method
- PCFG