Graduate Preliminary Exam: Spring 2014

Zhengyang Song

April 29, 2017

1 Longest palindrome subsequence

Solution: Let $dp[i][j]$ be the length of maximum subsequence from s[i...j]. Then we have

$$dp[i][i] = 1$$

$$dp[i][j] = 0, \forall i > j$$

$$dp[i][j] = max(dp[i+1][j], dp[i][j-1]) \text{ if } s[i] \neq s[j]$$

$$dp[i][j] = max(dp[i+1][j], dp[i][j-1], dp[i+1][j-1] + 2) \text{ if } s[i] == s[j]$$

Note dp array is in size $O(n^2)$, thus our algorithm running time.

2 Monge Matrix

$$A[i,j] + A[k,l] \leq A[i,l] + A[k,j], \forall i < k \text{ and } j < l$$

$$\begin{pmatrix} 10 & 9 & 12 & 10 \\ 11 & 10 & 13 & 10 \\ 9 & 8 & 0 & 7 \\ 11 & 10 & 11 & 8 \end{pmatrix}$$

Let $f(i)$ be the column index of the leftmost minimum element of row i.

1. Show that $f(1) \leq f(2) \leq \cdots \leq f(m)$.
2. Given an $m \times n$ Monge matrix, design an algorithm that computes $f(1), \ldots, f(m)$ in $O(m + n \log m)$ time.

Solution:

1. For $i < j$, we want to prove $f(i) \leq f(j)$. Suppose otherwise, we have $f(i) = s > f(j) = t$. Then

$$A[i,s] < A[i,t]$$

$$A[j,t] \leq A[j,s]$$

$$A[i,s] + A[j,t] < A[i,t] + A[j,s]$$

   It contradicts with the definition of Monge matrix.

2. We use divide-and-conquer. We first find $f(\lfloor m/2 \rfloor)$ in time $O(n)$. Then based on the above conclusion, we can safely drop the bottom left quarter and top right quarter. Do the same operation for the top left quarter and bottom right quarter. Note that this procedure will end in $O(\log m)$ rounds. In each round, the total time cost for all the subtasks are $O(n)$. Furthermore, output all the results will cost time $O(m)$. That leads us to the time complexity of $O(m + n \log m)$.

3

An interval graph is undirected graph $G = (V, E)$ where vertices correspond to intervals on the real line, where each interval is specified by a leftmost value v_1 and a rightmost value v_2. Two vertices in G are connected iff the corresponding intervals overlap. Let G be an interval graph whose corresponding intervals are provided. Give a polynomial time algorithm to find a maximum independent set in G. You need to prove the correctness of your algorithm. (An independent set in a graph is a subset of vertices such that no two vertices in the subset are joined by an edge.)

Solution: This problem is equivalent to find the maximum set of intervals with no intersection. We can use greedy algorithm. First sort the intervals using the rightmost value as key. Then go through the sorted list, each time we select the interval with the minimal rightmost value that has no intersection with the already selected ones. We show the correctness as follows.

Suppose there is another set of intervals S' that is different from the result of our algorithm S. Then after sorting S', we can always find the first interval i' in S' that is different from i in S. We simply replace i' with i, note that this will not violent the independent property. We can always do this until the last element of S'. Thus we have $|S'| \leq |S|$

4

Show that finding a min-cost matching with exactly k edges in a bipartite graph $G(U, V; E)$ (with positive edges) can be solved in polynomial time. Note that $|U|$ may not be the same as $|V|$. You can assume that the min-cost perfect matching problem in a bipartite graph can be solved in polynomial time.

Solution: We can turn this into a minimum-cost maximum-flow probelm. Add a source s and a sink t, together with a s' and t', where there is an edge with capacity k from s to s', and an edge with capacity 1 from s' to all the vertices in U. The same is with t' and t. The edges in E are all assigned with a capacity 1. All newly added edges are assigned with a cost 0. Then we just need to compute the minimum-cost maximum-flow of this new graph G', where a polynomial algorithm exists.

5 Finding the maximum area polygon

Solution: We can use a dynamic programming method. Suppose the polygon is rooted at 0, then we let $dp[m][i]$ be the maximum area we can get for a m-gon with the largest index vertex i.

$$dp[m][i] = max(\{dp[m-1][j] + area(0, i, j) | j < i\})$$

Then the answer for root 0 is just

$$answer = max(\{dp[n][i] | 1 \leq i \leq n\})$$

Then we enumerate all the possible root position (rather than 0) in time $O(n)$. That completes our polynomial algorithm.

6 Search

When are the following search algorithm optimal (i.e., find a solution that satisfies the goal conditions while incurs minimum cost)? Please be precise about the conditions you define.

1. BFS
2. DFS
3. A*

Solution:

1. BFS is optimal if the path cost is a nondecreasing function of the depth of the node. The most common such scenario is that all actions have the same cost. (AIMA P82).
2. The advantage of DFS over BFS is in the space complexity. (AIMA P86) So when there is only one unique goal with a deep path, DFS is optimal.
3. the tree-search version of A^* is optimal if $h(n)$ is admissible, while the graph-search version is optimal if $h(n)$ is consistent (AIMA P95).

Or refer to AIMA P108 for a summary of all kinds of search algorithms.

7 Probability theory and probabilistic reasoning

1. **Solution:**

$$P(young|use) = \frac{P(young)P(use|young)}{P(use)} = \frac{0.8}{0.8 + 0.2 + 0.4 \times 2}$$

2. A Bayesian Network is a DAG to represent the conditional dependencies for a set of random variable. Given n random variables, how many possible Bayesian Networks could be constructed to represent their conditional dependencies?

   Solution: The question is how many different DAGs can be formed using n nodes.

$$G(n) = \sum_{m=1}^{n} \sum_{n_1+\cdots+n_m=m, n_i \geq 1} \prod_{i=1}^{m-1} (2^{[i]} - 2^{[i-1]})^{n_{i+1}}$$

   where $[k] = \sum_{h=1}^{k} n_h$

3. **Solution:** No. Since D is dependent on C, and C is dependent on A. Image $P(C = true|A = true) = 1$ and $P(D = true|C = ture) = 1$, then $P(D = true|A = true) = 1$. The fact that no edge from A to D only means that A does not directly influence D.

$$P(E|A) \quad = \quad P(B|A)P(C|A, B)P(E|C)$$

4. Are small Markov blankets better when performing inference using Gibbs sampling? Briefly explain in one or two sentences why or why not.

   Solution: Yes. Because when Markov blanket is large, the inference won't change much.

8 General machine learning

1. In the curve fitting problem on a set of data points, what is the disadvantage of Nearest Neighbors approach compared to Least Squares method?

   Solution: Every time we need to start the computation from zero. It is computationally expensive to find the k nearest neighbors when the dataset is very large. Performance depends on the number of dimensions that we have.

   On the other hand, it is important how we compute the distance (there may be unrelated features). Also when the dimension is large, the sample data tends to be located at the corner, causing the distance metric meaningless. One solution is to use weighed features to compute the distance. (Credit to Tianqi Zhao)

2. True/False "When training a linear regression estimator, 10-fold cross-validation has smaller bias than 5-fold cross-validation"

   Solution: Yes. Bias is different from variance. The model we train as a part of the testing procedure, would be as close as possible to the one that we would get if we trained it on the entire dataset.

9 Maximum Likelihood (ML)

1. $$L(\theta) = \theta \cdot 3\theta \cdot (1-3\theta) \cdot (1-3\theta) = 3\theta^2(3\theta-1)^2$$

2. $$L(\theta) = \frac{1}{3}(3\theta(1-3\theta))^2 \leq \frac{1}{3}(\frac{3\theta+1-3\theta}{2})^4 = \frac{1}{48}$$

   where the equation sign is when

   $$3\theta = 1-3\theta \Rightarrow \theta = \frac{1}{6}$$

10 Probabilistic model learning

Given training data $x_1, \dots, x_n, \dots, x_N$, single densities $i \in 1, \dots, I$ where I is the total number of densities and unknown model parameter λ, estimate the model density $p(i|\lambda)$ in mixture density $p(x_n|\lambda) = \sum_i p(i|\lambda) \cdot p(x_n|i, \theta_i)$ using Expectation Maximization, according to the maximum likelihood criterion.

Solution:

$$p(x_n|\lambda) = \sum_i p(i|\lambda) \cdot p(x_n|i, \theta_i)$$

Here i is the hidden variable, and θ_i is model parameter. In the E step, we compute the probability that the data point x_j comes from density i

$$p_{ij} = P(I = i|x_j) = \alpha P(x_j|I = i, \theta_i)P(I = i|\lambda)$$

In the M step, we compute the θ_i maximize the probability that we get these data points.

$$E = \sum p_{ij} \log P(x_j|I = i, \theta)$$

We update the θ

$$\theta = \arg\max_\theta E$$

Then we update the density estimation.

$$P(I = i|\lambda) = \sum_j p_{ij}$$

11 Maximum a Posteriori (MAP)

$$P[p = 0.1] = 0.3, P[p = 0.3] = 0.7$$

Solution:

$$P(\theta = 0.1|R) = P(\theta = 0.1)P(R|\theta = 0.1) = 0.3 \times [0.1 \times (3 \times 0.1) \times (1 - 3 \times 0.1) \times (1 - 3 \times 0.1)]$$

$$P(\theta = 0.3|R) = P(\theta = 0.3)P(R|\theta = 0.3) = \dots$$

Then do a normalization.

12 Cross-Validation

What is the advantage of using a larger value, e.g. $k = 10$, rather than e.g. $k = 2$? Is there any advantage of using $k = 2$ rather than $k = 10$?

Solution: A larger k will introduce smaller bias, while a larger variance.

13 Clustering

1. Which of the two choices Single Linkage or Complete Linkage clustering will typically result in clusters with a small diameter? Why?

   Solution: In single-link clustering or single-linkage clustering , the similarity of two clusters is the similarity of their most similar members. In complete-link clustering or complete-linkage clustering , the similarity of two clusters is the similarity of their most dissimilar members.

   Complete Linkage. Since in Single Linkage clustering, more distant parts of the cluster and the clusters' overall structure are not taken into account

2. The algorithm's output can be visualized using a dendrogram. Explain in one or two sentences if there are any disadvantages of dendrograms on this dataset.

   Solution: The dendrograms may be complex, and the horizontal split line may be hard to recognize.

3. Explain in one or two sentences if there would be any advantage of instead using the k-means algorithm on this dataset.

   Solution: The time complexity and space complexity is lower.

14 SVM

x_{i1}	x_{i2}	y_i	α_i
1	4	1	2
1.5	7	1	0
2	6	1	0

1. What are the support vectors of this training data?

   Solution: where the $\alpha_i \neq 0$.

2. Assume that the decision surface bias is $b = -1.8$. Compute the weight vector w using the data in the table given above.

   Solution:

   $$w = \sum \alpha_i x_i y_i$$

3. Consider the points $(1, 4), (4.5, 3), (2, 2)$ and $(3, 1)$ from the table above. How are these points classified by the model? Compute the slack variables x_i for these points.

4. Use the SVM model to compute a classification for a test data point $(4, 4)$.

   Solution:

   $$y = w^T x + b$$

15 Perceptron and Kernel Methods

1. Does the Perceptron algorithm have any advantages over the regular SVM algorithm?

   Solution: Perceptrons can be trained online (i.e. their weights can be updated as new examples arrive one at a time).

   $$w \leftarrow w + y_i x_i$$

2. Under what circumstance can a machine learning algorithm be kernelized (i.e, adapted to work with kernels)? Can you find a way to do this for the Perceptron algorithm? Explain the changes you need to make to it and why these changes are acceptable. Also explain whether this algorithm has any disadvantages.

   Solution:
$$w \cdot x + b = \sum_{j \in E} y_j x_j \cdot x + b$$

We need to find a good kernel function.

16 Graphical Models

In this MRF model, the associated potential functions are defined by

$$\phi(x,y) = 2xy + x^3$$
$$\phi(x,z) = xz + z$$
$$\phi(y,z) = yz^3$$

1. Compute the posterior marginal probability $P(y|x=1)$

   Solution:
$$P(x,y,z) = \frac{1}{Z}\prod \phi(x,y)\phi(y,z)\phi(x,z)$$

$$\begin{aligned} Z &= \sum\prod \phi(x,y)\phi(y,z)\phi(x,z) \\ &= (2+1)(1+1)(1) + \cdots \end{aligned}$$

$$P(x,y) = \sum_z P(x,y,z)$$

$$P(x) = \sum_{y,z} P(x,y,z)$$

$$P(y|x=1) = \frac{P(x=1,y)}{P(x=1)}$$

2. In the following three directed graphical models, assume variable y has been observed:

$$x \rightarrow y \rightarrow z$$
$$x \rightarrow y \leftarrow z$$
$$x \leftarrow y \rightarrow z$$

   In the above three directed models, argue whether x and z are conditionally independent, given y.

   Solution: Independent:

$$P(x,z|y) = \frac{P(x,z,y)}{P(y)} = \frac{P(x)P(y|x)P(z|y)}{P(y)} = P(x|y)P(z|y)$$

   Not independent:

$$P(x,z|y) = \frac{P(x,y,z)}{P(y)} = \frac{P(x)P(z)P(y|x,z)}{P(y)} \neq P(x|y)P(z|y)$$

   Independent:

$$P(x,z|y) = \frac{P(x,y,z)}{P(y)} = \frac{P(y)P(x|y)P(z|y)}{P(y)} = P(x|y)P(z|y)$$

17 Ensemble learning

Write a rigorous derivation showing that that bagging decreases the mean squared error rate of a regression task, compared to the single regressor trained on D.

Solution: PRML P656.

18 Hidden Markov model

We can model $p(x_1^T) = \sum_{s_1^T} \prod_{t=1}^{T} p(x_t, s_t | x_1^{t-1}, s_1^{t-1})$.

1. Write down the model for $p(x_1^T)$ using dependencies only on the previous state (first-order Markov)
2. Write a generalized Hidden Markov Model that employs the forward algorithm for scoring.
3. What is the time complexity of the algorithm in direct calculation of (1) and forward algorithm of (2).

19 EM

Explain and prove the convergence of EM algorithm.

Solution:

- Expected complete data log likelihood is a lower bound.
- EM monotonically increases the observed data log likelihood.

定理 9.1 设$P(Y\mid\theta)$为观测数据的似然函数，$\theta^{(i)}$ $(i=1,2,\cdots)$为EM算法得到的参数估计序列，$P(Y\mid\theta^{(i)})$ $(i=1,2,\cdots)$为对应的似然函数序列，则$P(Y\mid\theta^{(i)})$是单调递增的，即

$$P(Y\mid\theta^{(i+1)})\geqslant P(Y\mid\theta^{(i)}) \tag{9.18}$$

证明 由于

$$P(Y\mid\theta)=\frac{P(Y,Z\mid\theta)}{P(Z\mid Y,\theta)}$$

取对数有

$$\log P(Y\mid\theta)=\log P(Y,Z\mid\theta)-\log P(Z\mid Y,\theta)$$

由式 (9.11)

$$Q(\theta,\theta^{(i)})=\sum_Z \log P(Y,Z\mid\theta)P(Z\mid Y,\theta^{(i)})$$

令

$$H(\theta,\theta^{(i)})=\sum_Z \log P(Z\mid Y,\theta)P(Z\mid Y,\theta^{(i)}) \tag{9.19}$$

于是对数似然函数可以写成

$$\log P(Y\mid\theta)=Q(\theta,\theta^{(i)})-H(\theta,\theta^{(i)}) \tag{9.20}$$

在式 (9.20) 中分别取θ为$\theta^{(i)}$和$\theta^{(i+1)}$并相减，有

$$\begin{aligned}&\log P(Y\mid\theta^{(i+1)})-\log P(Y\mid\theta^{(i)})\\&\quad=[Q(\theta^{(i+1)},\theta^{(i)})-Q(\theta^{(i)},\theta^{(i)})]-[H(\theta^{(i+1)},\theta^{(i)})-H(\theta^{(i)},\theta^{(i)})]\end{aligned} \tag{9.21}$$

为证式 (9.18)，只需证式 (9.21) 右端是非负的．式 (9.21) 右端的第 1 项，由于$\theta^{(i+1)}$使$Q(\theta,\theta^{(i)})$达到极大，所以有

$$Q(\theta^{(i+1)},\theta^{(i)})-Q(\theta^{(i)},\theta^{(i)})\geqslant 0 \tag{9.22}$$

其第 2 项，由式 (9.19) 可得：

$$\begin{aligned}&H(\theta^{(i+1)},\theta^{(i)})-H(\theta^{(i)},\theta^{(i)})\\&\quad=\sum_Z\left(\log\frac{P(Z\mid Y,\theta^{(i+1)})}{P(Z\mid Y,\theta^{(i)})}\right)P(Z\mid Y,\theta^{(i)})\\&\quad\leqslant\log\left(\sum_Z\frac{P(Z\mid Y,\theta^{(i+1)})}{P(Z\mid Y,\theta^{(i)})}P(Z\mid Y,\theta^{(i)})\right)\\&\quad=\log P(Z\mid Y,\theta^{(i+1)})=0\end{aligned} \tag{9.23}$$