CM146, Fall 2017
Problem Set 5: Boosting, Unsupervised learning
Due Dec 7, 2017, 11:59pm

Submission instructions

- Submit your solutions electronically on the course Gradescope site as PDF files.
- If you plan to typeset your solutions, please use the LaTeX solution template. If you must submit scanned handwritten solutions, please use a black pen on blank white paper and a high-quality scanner app.

1 AdaBoost [5 pts]

In the lecture on ensemble methods, we said that in iteration t, AdaBoost is picking (h_t, β_t) that minimizes the objective:

$$\begin{aligned}(h_t^*(\boldsymbol{x}), \beta_t^*) &= \underset{(h_t(\boldsymbol{x}),\beta_t)}{\arg\min} \sum_n w_t(n) e^{-y_n \beta_t h_t(\boldsymbol{x}_n)} \\ &= \underset{(h_t(\boldsymbol{x}),\beta_t)}{\arg\min} (e^{\beta_t} - e^{-\beta_t}) \sum_n w_t(n) \mathbb{I}[y_n \neq h_t(\boldsymbol{x}_n)] \\ &\qquad + e^{-\beta_t} \sum_n w_t(n)\end{aligned}$$

We define the weighted misclassification error at time t, ϵ_t to be $\epsilon_t = \sum_n w_t(n)\mathbb{I}[y_n \neq h_t(\boldsymbol{x}_n)]$. Also the weights are normalized so that $\sum_n w_t(n) = 1$.

(a) Take the derivative of the above objective function with respect to β_t and set it to zero to solve for β_t and obtain the update for β_t.

(b) Suppose the training set is linearly separable, and we use a hard-margin linear support vector machine (no slack) as a base classifier. In the first boosting iteration, what would the resulting β_1 be?

2 K-means for single dimensional data [5 pts]

In this problem, we will work through K-means for a single dimensional data.

(a) Consider the case where $K = 3$ and we have 4 data points $x_1 = 1, x_2 = 2, x_3 = 5, x_4 = 7$. What is the optimal clustering for this data ? What is the corresponding value of the objective ?

Parts of this assignment are adapted from course material by Jenna Wiens (UMich) and Tommi Jaakola (MIT).

(b) One might be tempted to think that Lloyd's algorithm is guaranteed to converge to the global minimum when $d = 1$. Show that there exists a suboptimal cluster assignment (*i.e.*, initialization) for the data in the above part that Lloyd's algorithm will not be able to improve (to get full credit, you need to show the assignment, show why it is suboptimal *and* explain why it will not be improved).

3 Hidden Markov Models [5 pts]

Consider a Hidden Markov Model with two hidden states, $\{1, 2\}$, and two possible output symbols, $\{A, B\}$. The initial state probabilities are

$$\pi_1 = P(q_1 = 1) = 0.49 \quad \text{and} \quad \pi_2 = P(q_1 = 2) = 0.51,$$

the state transition probabilities are

$$q_{11} = P(q_{t+1} = 1 | q_t = 1) = 1 \quad \text{and} \quad q_{12} = P(q_{t+1} = 1 | q_t = 2) = 1,$$

and the output probabilities are

$$e_1(A) = P(O_t = A | q_t = 1) = 0.99 \quad \text{and} \quad e_2(B) = P(O_t = B | q_t = 2) = 0.51.$$

Throughout this problem, make sure to show your work to receive full credit.

(a) There are two unspecified transition probabilities and two unspecified output probabilities. What are the missing probabilities, and what are their values?

(b) What is the most frequent output symbol (A or B) to appear in the first position of sequences generated from this HMM?

(c) What is the sequence of three output symbols that has the highest probability of being generated from this HMM model?