# COMS 4771 Spring 2017 Homework 5

Jin Tack Lim, jl4312

Problem 1.
(a)

- Initialization

  random $p^j(1), ..., p^j(D)$ where j = 1, ..., K

  $\pi_j = \frac{1}{K}$

- Expectation

  The probability that n-th document belongs to the class i. For t-th iteration, the superscript t for each of $\tau, \pi, p^j(1), ..., p^j(D)$ are omitted.

$$\tau_{n,i} = \frac{\pi_i \Pi_{d=1}^{D} p^i(d)^{x_n(d)}}{\sum_{j=1}^{K} \pi_j \Pi_{d=1}^{D} p^j(d)^{x_n(d)}}$$

- Maximization Update $p^i(1), ..., p^i(D)$ and $\pi_i$ where i = 1, ..., K

$$\pi_i^{(t+1)} = \frac{\sum_{q=1}^{n} \tau_{q,i}^{(t)}}{n}$$

$$p^{i(t+1)}(d) = \frac{\sum_{q=1}^{n} \tau_{i,q} x(d)}{\sum_{q=1}^{n} \tau_{i,q} l_q}$$

  where d=1, ..., D, q = 1, ..., n

(b)
The marginal distribution of $x_n$ is

$$\frac{\hat{x_n}}{l_n}$$

This can be interpreted as the binomial distribution for each dimensions. So if $l_n$ is very large and $\frac{\hat{x_n}}{l_n}$ is very small, it can be approximated to the Poisson distribution, where $\lambda = \hat{x_n}$.
(c)

1. Similarity

   - Both of them can reduce the dimensionality.

2. Difference

   - PCA gives vectors in order such that the variance along the vector is high. (e.g. the given data has the highest variance along the first vector.) For (b), there's no guarantee.

- PCA gives vectors that are orthogonal to each other, since they are all eigenvectors. For (b), there's no guarantee.

(d)

(b)

Consider the case in Figure **??**. There are three cases: C1, C2 and C3. Using 'one vs. one', we can draw three lines (3*2/2), which classify a sample to one of two classes for each pair of classes. The whole space is divided into seven regions. Out of seven, six regions are clearly classified into a specific class: R6 and R7 are C1, R2 and R3 are C2, R4 and R5 are C3. However R1 can't be classified because there is no single winner: 1 vote for each classes.

So, this example showed that 'one vs. one' also has some drawbacks.

(c)
1. The primal and quadratic problem:

$$\arg\min_{w,b} \frac{1}{2}\sum_{m=1}^{k}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \quad \text{such that}$$

$$y_i(w_m^T x_i + b) - 1 + \xi_i \geq 0 \quad \text{for each dimension m (from 1 to k)}$$

With Lagrange $\alpha$:

$$\arg\min_{w,b} \frac{1}{2}\sum_{m=1}^{k}\|w\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\sum_{m=1}^{k}\alpha_i^m(y_i(w_m^T x_i + b) - 1 + \xi_i) - \sum_{i=1}^{N}\beta_i\xi_i$$

all $\alpha$ and $\beta \geq 0$

2. The dual problem:

$$\arg\max_{\alpha}\sum_{i,m}\alpha_i^m - \frac{1}{2}\sum_{i,j,m}\alpha_i^m\alpha_j^m y_i y_j x_i^T x_j$$

$$\text{subject to} \quad 0 \leq \sum_{m}\alpha_i^m \leq C \quad \text{and} \sum_{i=1}^{N}\sum_{m=1}^{k}\alpha_i^m y_i = 0$$

Problem 2.

(a)

$$\sum_{i=1}^{k} p_i \log p_i$$

This is the Entropy without the negative sign.
This is how I got the expectation.
(1) When the value of $y_i$ is $j$, then its contribution is $\log p_j$.
(2) The probability that $y_i$ becomes $j$ is $p_j$.
(3) Therefore the expectation is the sum of $p_j \log p_j$ over all possible j (1,...,k)

(b)

$$\sum_{i=1}^{k} p_i \log q_i$$

Problem 3.

a) The marginal distribution is the binomial distribution. More specifically, the probablity of being weight k is

$$\binom{n}{k}\Big(\frac{1}{n}\Big)^k\Big(\frac{n-1}{n}\Big)^{n-k}$$

The correlation of each weight is, intuitively, if one of the weight is k, then the other weight should be equal to or less than n-k. In other words, the distribution of the former is Bin(n, p) and the latter is Bin(n-k, p)

b) This is a hard problem, so I read some papers[1][2][3] and borrowed their idea.

Step 1. Show that the error $\epsilon$ from the booster(i.e. Adaboost) is bounded by the product of $Z_t$.

$$\epsilon = \frac{1}{m} * \text{number of mispredicted points} \leq \prod_t Z_t$$

Step 2. Show that the product of Z is bounded by the expression with T and $\gamma$.

$$\prod_t Z_t \leq \exp(-2T\gamma^2)$$

Step 3. Device the given form in the problem from the step 1 and 2.

$$T = -\frac{\gamma^2}{C\log\epsilon}$$

Here are the detailed steps.

Step 1.

$$\epsilon = \frac{1}{m} * \text{number of mispredicted points} \leq \frac{1}{m}\sum_i \exp(-y_i f(x_i)) = \prod_t Z_t$$

The inequality comes from the fact that $\exp(-y_i f(x_i)) \geq 1$ if the data point is mispredicted. The equality comes by unraveling the recursive definition of $D_t$

Step 2.

This step is way too much complicated for me to digest.

Step 3.

By combining the result from step 1 and step 2, we can derive this form. C in this case is $\frac{1}{2}$

$$T \leq -\frac{\log\epsilon}{2\gamma^2}$$

References

[1] A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, Freund and Schapire, http://www.face-rec.org/algorithms/Boosting-Ensemble/decision-theoretic_generalization.pdf

[2] Improved Boosting Algorithms Using Confidence-rated Predictions, Schapire and Singer, http://web.cs.iastate.edu/ honavar/singer99improved.pdf

[3] The Boosting Approach to Machine Learning An Overview, Schapire, https://www.cs.princeton.edu/courses/
survey.pdf