Machine Learning for Statistics (Spring 2017) Homework #3

Minxia Ji - mji4@fordham.edu
Discussants: Google

April 19, 2017

Problem 1

When K = {1,2,3,4,5}, I got results below.When k = 3, the log likelihood of training data is maximized.When k = 1, the log likelihood of testing data is maximized. My friend is most likely to have 1 rigged coin in his bag and the probability of this rigged coin appears head is 0.49096.

Filter

	K=1	K=2	K=3	K=4	K=5
1st iteration	-17324.59	-1.523117e+04	-1.468526e+04	-14756.54294	-14823.82339
2nd iteration	-17324.59	-1.523139e+04	-1.468526e+04	-14778.19808	-14819.66846
3rd iteration	-17324.59	-1.523116e+04	-1.468526e+04	-14785.31690	-14870.73093
4th iteration	-17324.59	-1.523302e+04	-1.468526e+04	-14786.87409	-14780.89827
5th iteration	-17324.59	-1.523131e+04	-1.468526e+04	-14685.33656	-14821.91643
6th iteration	-17324.59	-1.523116e+04	-1.468526e+04	-14698.95199	-14806.14675
mean	-17324.59	-1.523153e+04	-1.468526e+04	-14748.53676	-14820.53070
standard deviation	0.00	7.335118e-01	7.452531e-04	45.20915	29.36391

(a) Training

Filter

	K=1	K=2	K=3	K=4	K=5
1st iteration	-17330.31	-20308.00067	-2.183374e+04	-21906.2167	-21987.15596
2nd iteration	-17330.31	-20319.61764	-2.183392e+04	-21927.2712	-21968.30618
3rd iteration	-17330.31	-20307.00360	-2.183376e+04	-21933.9956	-22019.95145
4th iteration	-17330.31	-20387.25944	-2.183379e+04	-21935.8751	-21939.43482
5th iteration	-17330.31	-20316.62110	-2.183384e+04	-21833.9791	-21978.96679
6th iteration	-17330.31	-20306.72937	-2.183395e+04	-21847.9661	-21958.86901
mean	-17330.31	-20324.20530	-2.183383e+04	-21897.5506	-21975.44737
standard deviation	0.00	31.36331	8.517493e-02	45.2888	27.39647

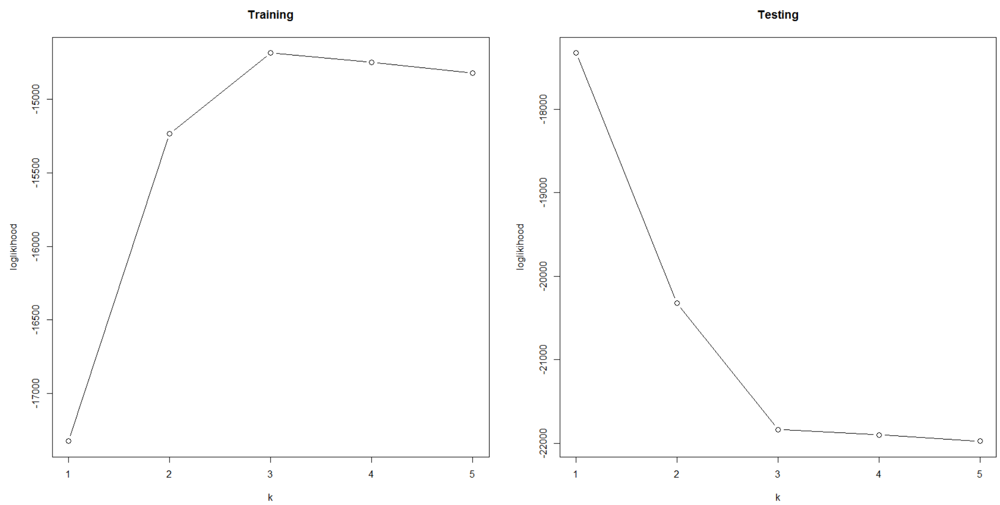
(b) Testing

K \ iterations	1st	2nd	3rd	4th	5th	6th
K=1	1	1	1	1	1	1
K=2	0.5144811	0.5206652	0.5136343	0.4863231	0.5136634	0.5135996
	0.4855189	0.4793348	0.4863657	0.5136769	0.4863366	0.4864004
K=3	0.3468591	0.3418439	0.3418523	0.3112076	0.3469503	0.3111816
	0.3418444	0.3113413	0.3115048	0.3469447	0.3112020	0.3418458
	0.3112965	0.3468148	0.3466429	0.3418477	0.3418477	0.3469725
K=4	0.3410804	0.04138035	0.34684825	0.2469981	0.341832222	8.042324e-05
	0.3468379	0.34183363	0.08874939	0.1000004	0.002509074	3.114442e-01
	0.2009852	0.31124042	0.34117513	0.3417814	0.344184680	3.418347e-01
	0.1110965	0.30554560	0.22322723	0.3112202	0.311474024	3.466406e-01
K=5	0.3281392	0.3410546	0.001287748	0.1540871	0.3077494	0.30736385
	0.3028902	0.2067838	0.307056806	0.3442491	0.2053457	0.34423729
	0.0227422	0.1161371	0.145391499	0.2169524	0.1427384	0.17074442
	0.1802357	0.1400722	0.202608111	0.1311702	0.1822338	0.12888788
	0.1659926	0.1959523	0.343655836	0.1535413	0.1619327	0.04876657

(c) π

K \ iterations	1st	2nd	3rd	4th	5th	6th
K=1	0.48932	0.48932	0.48932	0.48932	0.48932	0.48932
K=2	0.6970746	0.6944880	0.6974292	0.2695243	0.6974170	0.6974437
	0.2691724	0.2664615	0.2695430	0.6974114	0.2695302	0.2695582
K=3	0.2052388	0.4991726	0.4990179	0.7950288	0.2052752	0.7950401
	0.4992138	0.7949704	0.7948991	0.2052730	0.7950312	0.4993192
	0.7949900	0.2052212	0.2051526	0.4992944	0.4992995	0.2052841
K=4	0.4988712	0.2012988	0.2052356	0.2072607	0.4990555	0.2048179
	0.2052316	0.4992698	0.8124018	0.2004480	0.2051149	0.7949254
	0.7858463	0.7950145	0.4989199	0.4993100	0.2051735	0.4990818
	0.8104598	0.2058034	0.7876063	0.7950232	0.7949124	0.2051840
K=5	0.4944953	0.4988758	0.8055155	0.7964056	0.7963625	0.7964828
	0.7976185	0.2039720	0.7961154	0.2043031	0.4735876	0.2043031
	0.6368873	0.8099572	0.4623193	0.4754535	0.5372404	0.4710284
	0.2000441	0.2071109	0.5254752	0.5400767	0.2028734	0.5154181
	0.2104100	0.7855049	0.2041206	0.7964003	0.2058601	0.5603146

(d) μ



(e) Train vs Test

Problem 2

1.

$$f(x,m,k) = \frac{m!}{x_1!...x_n!}\mu^{x_1!...x_n!} = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(\sum_i x_i + 1)} \prod_{i=1}^{n} \mu_i^{x_i}$$

Let $mu_j^{x_n} = \prod_{l=1}^{M} \mu_j(l)^{x_n(l)}$.The E-step is written as follow.

$$T_{n,j} = p(Z_n = j|x_n, \theta) = \frac{\pi_j p(x_n|\mu_j)}{\sum_{i=1}^{K} \pi_i p(x_n|\mu_j)} = \frac{\pi_j \mu_j^{x_n}}{\sum_{i=1}^{K} \pi_i \mu_{x_n}}$$

In the M-steo we seek to maximize the parameters holding above fixed:

$$\Theta = \theta \sum_{n=1}^{N} \sum_{j=1}^{K} \times \log(\frac{p(x_n, z = j|\theta)}{T_{n,j}})$$

Now that$p(x_n, z = j|\theta) = p(x_n|z = j, \theta)p(z_n = j|\theta) = \pi_j \mu_j^{x_n}$ by the probability chain rule.

$$\mu_1, ..., \mu_k, \pi_1, ..., \pi_k \sum_{n=1}^{N} \sum_{j=1}^{K} T_{n,j}(\log(\mu_j^{x_n})) + \log(\pi_j))$$

subject to

$$\sum_{l=1}^{M} \mu_j(l) = 1 \qquad \forall j \in \{1, ..., K\}$$

$$\sum_{i=1}^{K} \pi_i = 1$$

Then we can write Lagrangian primal form as:

$$L(\mu,\pi,\alpha,\beta) = \sum_{n=1}^{N}\sum_{j=1}^{K} T_{n,j}(\log(\mu_j^{x_n}) + \log(\pi_j)) - \alpha(\sum_{j=1}^{K}\pi_j - 1) - \sum_{j=1}^{K}\beta_j(\sum_{l=1}^{M}\mu_j(l) - 1)$$

Differentiate with respect to parameters of interest starting with pi_j:

$$\frac{\partial L}{\partial \pi_j} = \sum_{n=1}^{N}\frac{T_{n,j}}{pi_j} - \alpha = 0$$

$$\pi_j = \frac{\sum_{n=1}^{N} Tn,j}{\alpha}$$

Plug into the primal constraint:

$$\frac{\sum_{n=1}^{N} Tn,j}{\sum_{n=1}^{N}\sum_{j=1}^{K} Tn,i}$$

Not surprisingly we find the MLE of the mixing components pi_j is the sample average of Tn,j

$$\pi_j = \frac{1}{N}\sum_{n=1}^{N} Tn,j$$

Repeating this process for $\mu_j(l)$:

$$\frac{\partial L}{\partial \mu_j(l)} = \sum_{n=1}^{N} T_{n,j}x_n(l)\frac{1}{\mu_j(l)} - \beta_j = 0$$

$$\mu_j(l) = \frac{1}{\beta_j}\sum_{n=1}^{N} T_{n,j}x_n(l)$$

Plug into the constraint on :

$$\mu_j(l) = \sum_{n=1}^{N}\frac{T_{n,j}}{\sum_{n'=1}^{N} T_{n,j}}x_n(l)$$

2.The marginal distribution of xn() for each word is Binomial. By the Poisson limit theorem as $l_n \to \infty$ and $x_n/l_n \to 0$:

$$\binom{n}{k}p^k(1-p)^{n-k} \approx e^{-\lambda}\frac{\lambda^k}{k!}$$

3. PCA assumes that the data is generated by a multivariate Gaussian distribution parametrized by its first two moments μ and Σ. The spectral decomposition gives a projection or linear transformation Λ that maximizes the variance of a linear combination of the original variables.

$$\Sigma = \Lambda \times D \times \Lambda^T$$

Rather than seeking vectors that are linear combinatons of the original variables, we are asked to find variables that can best approximate the document vectors:$\hat{x}_n = \sum_{\delta}^{d} u_n(\delta)v_\delta$

where both $u_n(\delta)$ and v_δ are hidden. This bears more similarity to factor analysis than principal component analysis. Also note that for the Poisson distribution whose first and second centered moment are λ, a projection that maximizes the variance of a linear combination of variables cannot be obtained via a spectral decomposition of the covariance matrix and the EM algorithm is needed.

4. We are given that the marginal distribution $\hat{x}_n(\delta) \backsim Bin(l_n, \frac{\hat{x}_n}{l_n})$ which we shall approximate with a Poisson distribution.

$$\hat{x}_n = \sum_{\delta}^{d} u_n(\delta) v_\delta$$

Denote the Poisson parameter $\lambda = \sum_{\delta}^{d} u_n(\delta) v_\delta$

$$P(x_n(\delta)|u_n, v) = \frac{\lambda^{x_n(\delta)} e^{-\lambda}}{x_n(\delta)!}$$

Writing the expected complete log likelihood and taking sample averages as maximizers of the suffi- cient statistics yields the following.

E-step:

$$T_{n,j} = \frac{\pi_i \sum_{\delta=1}^{d} (\frac{u_n(\delta) v_\delta}{l_n})}{\sum_j \pi_j \sum_{\delta=1}^{d} (\frac{u_n(\delta) v_\delta}{l_n})}$$

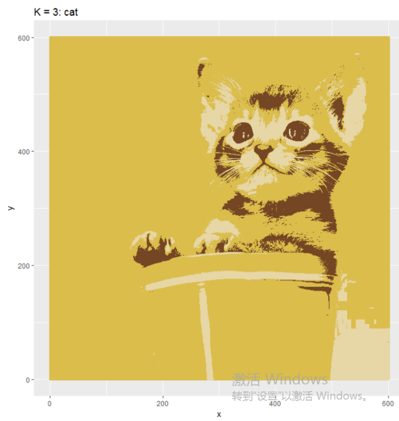
M-step:

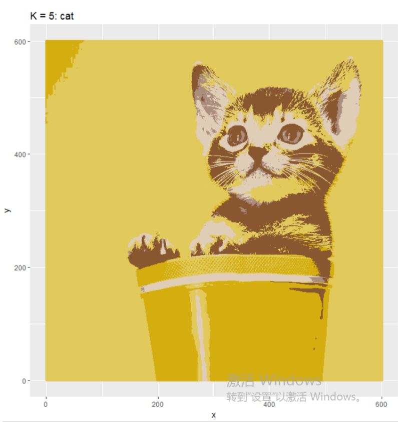
$$\pi_i = \frac{\sum_{n=1}^{N} T_{n,j}}{N}$$

$$u_n(\delta) = \frac{\sum_{n=1}^{N} v_\delta T_{n,i}}{\sum_{\delta'=1}^{d} \sum_{n=1}^{N} v'_\delta T_{n,i}}$$

Problem 3

Plot original cat picture.Then apply my function to the cat picture with k = 3 or 5.Original picture and changed pictures are below.





Problem 4

1.For sigmoid function, we have $y_k = g(a_k) = \frac{1}{1+exp(-a_k)}$ and take the derivative of y_k with respect to the activation a_k, we get $y_k' = y_k(1-y_k)$.

$$\frac{\partial E}{\partial a_k} = \frac{\partial}{\partial a_k}\{-[t_k ln(y_k) + (1-t_k)ln(1-y_k)]\}$$

$$= -[t_k \frac{\partial ln(y_k)}{\partial y_k}\frac{\partial y_k}{\partial a_k} + (1-t_k)\frac{\partial ln(1-y_k)}{\partial(1-y_k)}\frac{\partial(1-y_k)}{\partial a_k}]$$

$$= -(\frac{t_k}{y_k}y_k(1-y_k) - \frac{(1-t_k)}{1-y_k}(1-y_k)y_k)$$

$$= -[t_k(1-y_k) - (1-t_k)y_k] = y_k - t_k$$

Therefore, proved the derivative of the error function with respect to the activation a_k is $y_k - t_k$

2.With interpretation $y_k(x,\theta) = p(t_k = 1|x)$

$$E(\theta) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{k_n} ln(y_k)$$

For a data set with sample size = N, the likelihood function can be written as:

$$P(T|w_1, ..., w_k) = \prod_{n=1}^{N}\prod_{k=1}^{K} y_n k^{t_n k}$$

Then the log likelihood function would be:

$$lnP(T|w_1, ..., w_k) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{k_n} ln y_k(x_n, w) = -E(\theta)$$

Therefore, minimizing the cross entropy function is equal to maximizing the likelihood for a multiclass neural network.