The Classification of Simulated Gene Expression Data Using LDA, KNN, etc.

October 1, 2019

```
> set.e$train
              e1          e2          e3           e4           e5
1   -0.869215856 -0.23562203  0.433777067  5.537948e-01  1.063533e-02
2    0.556755587  0.98388789  0.411870967 -2.808091e-01 -2.318547e-01
3   -2.820701032  0.66934749  0.122762361 -5.452579e-01  2.245622e-01
4   -1.025282104  0.19911721  0.827533058  8.093052e-02 -2.422870e-01
5   -0.593736680 -1.42365411  0.649620527  1.639880e-01 -5.136332e-01
6   -1.092823476  0.44827997 -0.026507738 -5.537899e-01 -1.683859e-01
7   -1.186541380  0.66016536  0.543837469 -3.542383e-02  5.154803e-01
8   -3.024055210 -0.24716041  0.343763698  9.113634e-02 -6.634528e-01
9   -3.789155918  0.99908339 -0.144872988 -8.804926e-02  3.332286e-01
10  -1.969398241 -1.13112182 -1.070065773 -1.809187e-01  1.896484e-01
```

- With simulated data curves, five FPC scores were calculated

$$\hat{\epsilon}'_{im} = \sum_{k=1}^{S}((\hat{X}_i(k) - \hat{\mu}'(k))\hat{\rho}'_m(k),\ m = 1, ..., 5,\ S = 18$$

- Last time, the classification performance of logistic regression and SVM was compared
- Two methods didn't show significant difference
- This time, classification performed with LDA, QDA, KNN, and neural net

LDA and QDA

- Let $f_k(X) = Pr(X = x|Y = k)$ denote the density function of X for an observation(x) that comes from the kth class(Y)

- Then, Bayes' Theorem states that,

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_k(x)}, \text{ where } \pi_k = Pr(Y = k)$$

- Linear Discriminant Analysis(LDA) assumes that $f_k(x)$ is normal pdf with same variance across all classes:

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{(x-\mu)'\Sigma^{-1}(x-\mu)}{2}\right)$$

- On the other hand, Quadratic Discriminant Analysis(QDA) assumes each class has its own variance:

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{(x-\mu)'\Sigma_k^{-1}(x-\mu)}{2}\right)$$

LDA and QDA

- $\delta_k(x)$ is defined by plugging $f_k(x)$ into $Pr(Y = k|X = x)$ and taking logarithm

$$\text{LDA} : \delta_k(x) = x'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + \log\pi_k$$

$$\text{QDA} : \delta_k(x) = -\frac{1}{2}(x - \mu_k)'\Sigma_k^{-1}(x - \mu_k) + \log\pi_k$$

- Discriminant analysis assigns an observation to the class for which δ_k is largest

- $\hat{\mu}$ and $\hat{\Sigma}$ are estimated by data

K Nearest Neighbor Classifiers / Neural Network

- K nearest neighbor classifiers:
  1) Given an observation x_0, find the k training points $x_{(r)}$, $r = 1, ..., k$ closest in distance to x_0

  2) Classify using majority vote among the k neighbors

- Neural Network:
  1) 5 input, 2 hidden layers(first 10, second 5), 1 output

  2) Activation function is logistic

- Classification Error rates

Table: FPCA logistic regression and SVM(linear)

No. of FPCs or base functions	Group 1 FPCA	SVM(linear)	Group 2 FPCA	SVM(linear)	overall FPCA	SVM(linear)
1	32.72 (8.41)	63.20 (16.53)	32.70 (8.31)	47.38 (4.97)	32.71 (5.26)	55.29 (6.86)
2	22.16 (6.65)	21.90 (7.15)	22.06 (6.15)	22.80 (6.81)	22.11 (4.33)	22.35 (4.22)
3	7.58 (4.58)	7.60 (4.60)	8.26 (5.34)	8.32 (5.02)	7.92 (3.35)	7.96 (3.27)
4	7.14 (4.14)	6.86 (4.19)	7.62 (5.10)	7.82 (4.92)	7.38 (3.11)	7.34 (2.98)
5	7.40 (4.07)	7.14 (4.02)	7.86 (5.26)	7.88 (5.22)	7.63 (3.06)	7.51 (3.10)

Table: LDA, QDA, KNN, and neural net

No. of FPCs	Group 1 LDA	QDA	Group 2 LDA	QDA	overall LDA	QDA
1	31.66 (8.22)	31.22 (9.55)	33.56 (8.62)	34.00 (10.42)	32.61 (5.80)	32.61 (5.68)
2	21.70 (7.54)	22.24 (7.32)	22.02 (6.74)	22.24 (6.96)	21.86 (4.44)	22.24 (4.60)
3	7.32 (4.78)	7.44 (4.58)	7.74 (4.10)	7.78 (3.77)	7.53 (2.85)	7.61 (2.75)
4	6.40 (3.77)	6.76 (3.89)	7.08 (3.71)	7.22 (3.89)	6.74 (2.47)	6.99 (2.52)
5	6.40 (3.76)	7.12 (4.05)	7.24 (3.65)	7.74 (4.04)	6.82 (2.38)	7.43 (2.64)

No. of FPCs	Group 1 KNN(k=11)	nnet(10,5)	Group 2 KNN	nnet	overall KNN	nnet
1	33.70 (10.52)	34.44 (14.18)	37.46 (11.37)	37.08 (14.06)	35.58 (6.20)	35.76 (6.58)
2	23.14 (8.42)	29.28 (8.63)	24.60 (7.86)	30.42 (8.77)	23.87 (5.11)	29.85 (5.68)
3	12.12 (5.84)	11.24 (6.08)	12.80 (6.22)	11.78 (5.92)	12.46 (3.83)	11.51 (4.15)
4	11.66 (5.98)	9.44 (4.90)	12.32 (5.84)	10.20 (5.92)	11.99 (3.76)	9.82 (3.89)
5	11.74 (6.06)	10.02 (5.08)	12.32 (5.93)	9.46 (5.38)	12.03 (3.89)	9.74 (3.56)