CS 189: Introduction to Machine Learning

Fall 2017

•

Homework 8

•

Solutions by

Jinhong Du

3033483677

Question 1

(a)

Jinhong Du
jaydu@berkeley.edu

(b)

I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up. Jinhong Du

Question 2

(a)

$\because$

$$\mathbb{P}(w_1) = \mathbb{P}(w_2) = \frac{1}{2}$$

$\therefore$

$$\begin{aligned}\mathbb{P}(w_i|x) &= \frac{\mathbb{P}(x|w_i)\mathbb{P}(w_i)}{\mathbb{P}(x)} \\ &= \frac{\mathbb{P}(x|w_i)\mathbb{P}(w_i)}{\mathbb{P}(x|w_1)\mathbb{P}(w_1) + \mathbb{P}(x|w_2)\mathbb{P}(w_2)} \\ &= \frac{\mathbb{P}(x|w_i)}{\mathbb{P}(x|w_1) + \mathbb{P}(x|w_2)}\end{aligned}$$

$\therefore$ the optimal decision boundary is

$$\mathbb{P}(w_1|x) = \mathbb{P}(w_2|x)$$

or

$$\mathbb{P}(x|w_1) = \mathbb{P}(x|w_2)$$

$\because$

$$\mathbb{P}(x|w_i) \sim N(\mu_i, \sigma^2)$$

$\therefore$ the optimal decision boundary is

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu_1)^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu_2)^2}$$

i.e.

$$(x-\mu_1)^2 = (x-\mu_2)^2$$

i.e.

$$x = \frac{\mu_1 + \mu_2}{2}$$

The decision rule is :
If $x < \dfrac{\mu_1 + \mu_2}{2}$, i.e. $\mathbb{P}(x|w_1) > \mathbb{P}(x|w_2)$, i.e. $\mathbb{P}(w_1|x) > \mathbb{P}(w_2|x)$, then x belongs to class w_1;
If $x > \dfrac{\mu_1 + \mu_2}{2}$, i.e. $\mathbb{P}(x|w_1) < \mathbb{P}(x|w_2)$, i.e. $\mathbb{P}(w_1|x) < \mathbb{P}(w_2|x)$, then x belongs to class w_2.

(b)

The probability of misclassification (error rate) associated with this decision rule is

$$
\begin{aligned}
P_e &= \mathbb{P}(\text{misclassified as } w_1|w_2)\mathbb{P}(w_2) + \mathbb{P}(\text{misclassified as } w_2|w_1)\mathbb{P}(w_1) \\
&= \frac{1}{2}\mathbb{P}(\{x : x > \frac{\mu_1+\mu_2}{2}\}|w_2) + \frac{1}{2}\mathbb{P}(\{x : x < \frac{\mu_1+\mu_2}{2}\}|w_2) \\
&= \frac{1}{2}\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\int_{\frac{\mu_1+\mu_2}{2}}^{\infty} e^{-\frac{1}{2\sigma^2}(x-\mu_2)^2}\mathrm{d}x + \frac{1}{2}\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\frac{\mu_1+\mu_2}{2}} e^{-\frac{1}{2\sigma^2}(x-\mu_1)^2}\mathrm{d}x \\
&\overset{\textit{symmetric}}{=\!=\!=\!=} \frac{1}{\sqrt{2\pi\sigma^2}}\int_{\frac{\mu_1+\mu_2}{2}}^{\infty} e^{-\frac{1}{2\sigma^2}(x-\mu_2)^2}\mathrm{d}x \\
&\overset{z=\frac{x-\mu_2}{\sigma}}{=\!=\!=\!=} \frac{1}{\sqrt{2\pi}}\int_{a}^{\infty} e^{-\frac{1}{2}z^2}\mathrm{d}z
\end{aligned}
$$

where $a = \dfrac{\mu_2-\mu_1}{2\sigma}$.

(c)

$$
\begin{aligned}
\lim_{a\to\infty} P_e &= \lim_{a\to\infty}\frac{1}{\sqrt{2\pi}}\int_{a}^{\infty} e^{-\frac{1}{2}z^2}\mathrm{d}z \\
&= 0
\end{aligned}
$$

It shows that when μ_1 and μ_2 are very distinctly different, then the error rate will be small enough since it's more likely to make a good decision about which class x belongs to.

Question 3

(a)

Gaussian prior with smaller variance

For regression problem

$$\max \|wX - Y\|_2^2 + \lambda \|w\|_2^2$$

Suppose that $y_i = wx_i + \epsilon$ where $\epsilon \sim N(0, \sigma_1^2)$, and $w \sim N(0, \sigma_2^2 I)$,

$$\begin{aligned}
\hat{w}_{MAP} &= \arg\min_w \mathbb{P}(Y|X, \lambda, w)\mathbb{P}(w) \\
&= \arg\max_w \prod_{i=1}^n \frac{1}{\sigma_1^2}\|y_i - x_i w\|_2^2 + \frac{1}{\sigma_2^2}\|w\|_2^2 \\
&= \arg\max_w \|Y - Xw\|_2^2 + \frac{\sigma_1^2}{\sigma_2^2}\|w\|_2^2
\end{aligned}$$

i.e.

$$\lambda = \frac{\sigma_1^2}{\sigma_2^2}$$

And we know that regularization shrinks w, i.e. decreases the variance.

(b)

TLS allows errors in X and y. OLS only allows errors in y.

(c)

$f_1(x) = \max\{-x, 0.1x\}$, $f_2(x) = x + \frac{x^2}{10}$, $f_4(x) = \frac{e^x + e^{-x}}{2} - 1$ are convex.

(d)

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

x and y seem to have strong negative linear correlation, i.e. $\rho < 0$ and $|\rho|$ is big,

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

and here $\sigma_x^2 = \sigma_y^2 = 1$

$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ has $\rho = -1$.

(e)

If assuming that increasing k is increasing features, then choose **Training Error** and **Bias** since the model will fit the data better and better.
If assuming that increasing k is increasing samples, then choose **Variance** since more data will decrease variance.
Since we don't know k outside the plot, so it can be **Validation Error**

(f)

If assuming that increasing k is increasing features, then choose **Variance** since the model tends to be overfitting.
Since we don't know k outside the plot, so it can be **Validation Error**

(g)

Validation Error and **Variance**

Question 4

(a)

Covariance matrix should be positive semi-definite.

$$\begin{vmatrix} 4 & a \\ a & 1 \end{vmatrix} = 4 - a^2 \geqslant 0$$

i.e.

$$-2 \leqslant a \leqslant 2$$

(b)

We should use 2 principal components to represent this data set since we can use a subspace - a xy plane to represent the line.

Question 5

(a)

$\because$

$$p(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\lambda) = \prod_{i=1}^{n} P(X_i = x_i|\lambda)$$

$$= \frac{\lambda^{\sum\limits_{i=1}^{n} x_i} e^{-n\lambda}}{\prod\limits_{i=1}^{n} x_i!}$$

$\therefore$

$$\ln p(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\lambda) = \sum_{i=1}^{n} x_i \ln\lambda - n\lambda - \sum_{i=1}^{n} \ln(x_i!)$$

(b)

$\because \quad x_i \geqslant 0$

$\therefore$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}[-\ln p(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\lambda)] = -\frac{1}{\lambda}\sum_{i=1}^{n} x_i + n$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}[-\ln p(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\lambda)] = \frac{1}{\lambda^2}\sum_{i=1}^{n} x_i \geqslant 0$$

$\therefore \quad -\ln p(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\lambda)$ is a convex function

Let

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln p(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\lambda) = 0$$

we have

$$\hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\because \quad -\ln p(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\lambda)$ is a convex function

$\therefore \quad -\hat{\lambda}_{MLE}$ is the global minimum of $-\ln p(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\lambda)$, i.e. $\hat{\lambda}_{MLE}$ is the global maximum of $-\ln p(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\lambda)$

(c)

$$
\begin{aligned}
\max \mathbb{P}(\lambda | x_1, \cdots, x_n) &= \max \frac{\mathbb{P}(x_1, \cdots, x_n | \lambda) f(\lambda)}{\int_0^\infty \mathbb{P}(x_1, \cdots, x_n | \lambda) f(\lambda) \mathrm{d}\lambda} \\
&= \max \mathbb{P}(x_1, \cdots, x_n | \lambda) f(\lambda) \\
&= \max \ln \mathbb{P}(x_1, \cdots, x_n | \lambda) + \ln f(\lambda) \\
&= \max \sum_{i=1}^{n} x_i \ln \lambda - n\lambda - \sum_{i=1}^{n} \ln(x_i!) + \ln \alpha - \alpha\lambda \\
&= \max \sum_{i=1}^{n} x_i \ln \lambda - n\lambda - \alpha\lambda
\end{aligned}
$$

Let

$$
g(\lambda) = \sum_{i=1}^{n} x_i \ln \lambda - n\lambda - \alpha\lambda
$$

and

$$
\frac{\mathrm{d}}{\mathrm{d}\lambda} g(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - (n + \alpha) = 0
$$

we have

$$
\hat{\lambda}_{MAP} = \frac{1}{n + \alpha} \sum_{i=1}^{n} x_i
$$

$\because$

$$
\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} g(\lambda) = -\frac{1}{\lambda^2} \sum_{i=1}^{n} x_i \leqslant 0
$$

$\therefore$ $\hat{\lambda}_{MAP}$ is the global maximum of $g(\lambda)$

$\because$ when n is big enough,

$$
\hat{\lambda}_{MAP} \approx \frac{1}{n} \sum_{i=1}^{n} x_i = \hat{\lambda}_{MLE}
$$

Question 6

(a)

$$\hat{w'}_{OLS} = \arg\min \frac{1}{2}\|X'w' - y'\|_2^2$$
$$\hat{w'}_{OLS} = (X'^T X')^{-1} X'^T y'$$

$$\begin{aligned}
\frac{1}{2}\|X'w' - y'\|_2^2 &= \frac{1}{2}\left\|\begin{bmatrix} Xw' - y \\ ce_1^T w' \\ \vdots \\ ce_d^T w' \end{bmatrix}\right\|_2^2 \\
&= \frac{1}{2}\|Xw' - y\|_2^2 + \frac{1}{2}\sum_{i=1}^{d}\|ce_i^T w'\|_2^2 \\
&= \frac{1}{2}\|Xw' - y\|_2^2 + \frac{c^2}{2}\sum_{i=1}^{d} w_1'^2 \\
&= \frac{1}{2}\|Xw' - y\|_2^2 + \frac{c^2}{2}\|w'\|_2^2
\end{aligned}$$

$\therefore$

$$\lambda = c^2$$

(b)

We can formulate the problem as

$$\frac{1}{2}\left\|X'w' - y'\right\|_2^2$$

therefore, by choosing $\gamma = \dfrac{2}{\lambda_{\min}(X'^T X') + \lambda_{\max}(X'^T X')}$, the loss function will have geometric convergence.

$\because$

$$X'^T X' = X^T X + c^2 I_{d\times d}$$

$\therefore$

$$\lambda(X'^T X') = \lambda(X^T X) + c^2$$

i.e.

$$\gamma = \frac{2}{m + M + 2c^2} = \frac{2}{m + M + 2\lambda}$$

Question 7

(a)

Let Y_i, M_i and N_i be the ith column of Y, M and N respectively.

$\because$

$$Y = M + N$$

and $N_{ij} \overset{iid}{\sim} N(0,1)$

$\therefore \quad Y_i | M_i \sim N(0, I_{d\times d})$

$\because$

$$\begin{aligned}
\mathbb{P}(Y|M) &= \mathbb{P}(Y_1, \cdots, Y_d | M_1, \cdots, M_d) \\
&= \prod_{i=1}^{d} \mathbb{P}(Y_i|M_i) \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}\sum\limits_{i=1}^{d}(Y_i - M_i)^T(Y_i - M_i)} \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}\sum\limits_{i=1}^{d}\|Y_i - M_i\|_2^2} \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}\sum\limits_{i=1}^{d}\sum\limits_{j=1}^{d}(Y_{ji} - M_{ji})^2} \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}\|Y - M\|_F^2}
\end{aligned}$$

$\therefore$

$$\begin{aligned}
\arg\max_M \mathbb{P}(Y|M) &= \arg\max_M \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}\|Y-M\|_F^2} \\
&= \arg\min_M \|Y - M\|_F^2
\end{aligned}$$

(b)

$\because \quad Y$ is full rank

$\therefore \quad$ the singular value decomposion of Y is

$$Y = U\Sigma V^T$$

where $\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 \\ 0 & 0 & \cdots & \sigma_d \end{pmatrix}$ and $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_d$

Solution (cont.)

By the Eckart-Young Theorem, the closest $(d-1)$-rank matrix M to Y in F-norm is given by

$$\begin{aligned} M_{d-1} &= U \begin{pmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & \sigma_d \end{pmatrix} V^T \\ &= \sigma_d u_d v_d^T \end{aligned}$$

where $u_i,\ v_i$ is the ith column of $U,\ V$ respectively.

(c)

By the Eckart-Young Theorem, the closest k-rank matrix M to Y in F-norm is given by

$$\begin{aligned} M_k &= U \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_d \end{pmatrix} V^T \\ &= \sum_{i=k+1}^{d} \sigma_i u_i v_i^T \end{aligned}$$

where $u_i,\ v_i$ is the ith column of $U,\ V$ respectively.

Question 8

(a)

$$
\begin{aligned}
\mathbb{E}[\hat{X}\hat{Q}^T] &= \mathbb{E}[U^T\Sigma_{XX}^{-\frac{1}{2}}XQ^T\Sigma_{QQ}^{-\frac{1}{2}}V] \\
&= U^T\Sigma_{XX}^{-\frac{1}{2}}\mathbb{E}[XQ^T]\Sigma_{QQ}^{-\frac{1}{2}}V \\
&= U^T\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XQ}\Sigma_{QQ}^{-\frac{1}{2}}V \\
&= U^TCV \\
&= U^TU\Lambda V^TV \\
&= \Lambda \\
&= \Sigma_{\hat{X}\hat{Q}} \\
\Sigma_{\hat{X}\hat{X}} &= \mathbb{E}[\hat{X}\hat{X}^T] \\
&= \mathbb{E}[U^T\Sigma_{XX}^{-\frac{1}{2}}XX^T\Sigma_{XX}^{-\frac{1}{2}}U] \\
&= U^T\Sigma_{XX}^{-\frac{1}{2}}\mathbb{E}[XX^T]\Sigma_{XX}^{-\frac{1}{2}}U \\
&= U^T\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XX}\Sigma_{XX}^{-\frac{1}{2}}U \\
&= U^TU \\
&= I \\
\Sigma_{\hat{Q}\hat{Q}} &= \mathbb{E}[\hat{Q}\hat{Q}^T] \\
&= \mathbb{E}[V^T\Sigma_{QQ}^{-\frac{1}{2}}QQ^T\Sigma_{QQ}^{-\frac{1}{2}}V] \\
&= V^T\Sigma_{QQ}^{-\frac{1}{2}}\mathbb{E}[QQ^T]\Sigma_{QQ}^{-\frac{1}{2}}V \\
&= V^T\Sigma_{QQ}^{-\frac{1}{2}}\Sigma_{QQ}\Sigma_{QQ}^{-\frac{1}{2}}V \\
&= V^TV \\
&= I
\end{aligned}
$$

(b)

$$
\begin{aligned}
\mathbb{E}[(Y-w^T\hat{X})^2] &= \mathbb{E}[Y^2] + \mathbb{E}[(w^T\hat{X})^2] - 2\mathbb{E}[Yw^T\hat{X}] \\
&= \mathbb{E}[Y^2] + \mathbb{E}[w^T\hat{X}\hat{X}^Tw] - 2\mathbb{E}[w^T(Y\hat{X})] \\
&= \mathbb{E}[Y^2] + w^T\mathbb{E}[\hat{X}\hat{X}^T]w - 2w^T\mathbb{E}[Y\hat{X}] \\
&= \mathbb{E}[Y^2] + w^Tw - 2v\mathbb{E}[Y\hat{X}] \\
&= \mathbb{E}[Y^2] + \|w\|_2^2 - 2w^T\mathbb{E}[Y\hat{X}]
\end{aligned}
$$

(c)

$$
\begin{aligned}
&\arg\min_{w\in\mathbb{R}^d} \mathbb{E}[(Y - w^T\hat{X})^2] + \|w\|^2_{CCA} \\
&= \arg\min_{w\in\mathbb{R}^d} \|w\|_2^2 - 2w^T\mathbb{E}[Y\hat{X}] + w^T \begin{bmatrix} \frac{1-\lambda_1}{\lambda_1} & 0 & \cdots & 0 \\ 0 & \frac{1-\lambda_2}{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1-\lambda_d}{\lambda_d} \end{bmatrix} w \\
&= \arg\min_{w\in\mathbb{R}^d} \|w\|_2^2 - 2w^T\mathbb{E}[Y\hat{X}] + w^T\Lambda w \\
&= \arg\min_{w\in\mathbb{R}^d} g(w)
\end{aligned}
$$

Let

$$\frac{\mathrm{d}}{\mathrm{d}w} g(w) = 2w - 2\mathbb{E}[Y\hat{X}] + 2\Lambda w = 0$$

we have

$$\begin{bmatrix} \frac{1}{\lambda_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\lambda_d} \end{bmatrix} w = \mathbb{E}[Y\hat{X}]$$

i.e.

$$\begin{bmatrix} \frac{w_1}{\lambda_1} \\ \frac{w_2}{\lambda_2} \\ \vdots \\ \frac{w_d}{\lambda_d} \end{bmatrix} = \mathbb{E}[Y\hat{X}]$$

i.e.

$$\frac{w_i}{\lambda_i} = \mathbb{E}[Y(\hat{X})_i]$$

i.e.

$$\lambda_i = w_i\mathbb{E}[Y(\hat{X})_i]$$

(d)

$\because$

$$\hat{x}^j = U^T\Sigma_{XX}^{-\frac{1}{2}}x^j$$

$\therefore$

$$
\begin{aligned}
\mathbb{E}\hat{x}^j\hat{x}^{jT} &= U^T\Sigma_{XX}^{-\frac{1}{2}}\mathbb{E}[x^jx^{jT}]\Sigma_{XX}^{-\frac{1}{2}}U \\
&= U^T\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XX}^{-\frac{1}{2}}U \\
&= I
\end{aligned}
$$

$\therefore \quad Cov[(\hat{x}^j)_i, (\hat{x}^j)_k) = 0 \ (i \neq k)$ and $Var[(\hat{x}^j)_i, (\hat{x}^j)_i] = 1$

Solution (cont.)

$\therefore$

$$
\begin{aligned}
\mathbb{E}[\|\tilde{w}-\hat{w}\|_2^2] &= \mathbb{E}\|\tilde{w}-\mathbb{E}(\tilde{w})\|_2^2 \\
&= \sum_{i=1}^{d} \mathbb{E}[\tilde{w}_i-\hat{w}_i]^2 \\
&= \sum_{i=1}^{d} Var[\tilde{w}_i] \\
&= \sum_{i=1}^{d} Var\left[\frac{\lambda_i}{n}\sum_{j=1}^{n} y^j(\hat{x}^j)_i\right] \\
&= \sum_{i=1}^{d} \frac{\lambda_i^2}{n^2} Var\left[\sum_{j=1}^{n} y^j(\hat{x}^j)_i\right] \\
&= \sum_{i=1}^{d} \frac{\lambda_i^2}{n^2} Var\left[y^T(\hat{x})_i\right] \\
&= \sum_{i=1}^{d} \frac{\lambda_i^2}{n^2} \sum_{j=1}^{n} Var\left[y^j(\hat{x}^j)_i\right] \\
&\leqslant \sum_{i=1}^{d} \frac{\lambda_i^2}{n^2} \sum_{j=1}^{n} \mathbb{E}\left[y^j(\hat{x}^j)_i\right]^2 \\
&= \sum_{i=1}^{d} \frac{\lambda_i^2}{n^2} \sum_{j=1}^{n} \mathbb{E}\left[y^{j2}(\hat{x}^j)_i^2\right] \\
&\leqslant \sum_{i=1}^{d} \frac{\lambda_i^2}{n^2} \sum_{j=1}^{n} \mathbb{E}\left[(\hat{x}^j)_i^2\right] \\
&= \sum_{i=1}^{d} \frac{\lambda_i^2}{n}
\end{aligned}
$$

Question 9

Question How to do LDA with multiclass?

Solution

Suppose that each of N classes has a mean μ_i and the same covariance Σ. Then the scatter between class variability may be defined by the sample covariance of the class means

$$\Sigma_b = \frac{1}{N}\sum_{i=1}^{N}(\mu_i - \mu)(\mu_i - \mu)^T$$

where μ is the mean of the class means, μ_i is the mean of the ith class. The class separation in a direction $\vec{w}$ in this case will be given by

$$S = \frac{\vec{w}^T \Sigma_b \vec{w}}{\vec{w}^T \Sigma \vec{w}}$$

where Σ is the sample cavariance matrix of all data.This means that when $\vec{w}$ is an eigenvector of $\Sigma^{-1}\Sigma_b$ the separation will be equal to the corresponding eigenvalue.

If $\Sigma^{-1}\Sigma_b$ is diagonalizable, the variability between features will be contained in the subspace spanned by the eigenvectors corresponding to the $N-1$ largest eigenvalues (since Σ_b is of rank $N-1$ at most). These eigenvectors are primarily used in feature reduction, as in PCA.

We can use eigenvectors corresponding to the $N-1$ largest eigenvalues to split the data space into N classes.