Lecture 11, Feb 12, 2024

Gaussian Discriminant Analysis

- Consider a classification problem where we have classes $c \in C$, each having a prior π_c, jointly Gaussian distributed with a mean of $\boldsymbol{\mu}_c$ and a covariance of $\boldsymbol{\Sigma}_c$
- Gaussian discriminant analysis is a special case of hypothesis testing for this type of classification problem; given an observation of the vector $\boldsymbol{X}$, we would like to know which class it came from (i.e. which hypothesis is true)
- Consider the case where all the classes have the same covariance $\boldsymbol{\Sigma}$, so they differ only by their mean
  - The posterior is $P[y = c|\boldsymbol{x}] = \dfrac{f(\boldsymbol{x}|c)\pi_c}{\sum_{c'} f(\boldsymbol{x}|y = c')\pi_c'}$
  - The numerator becomes $\dfrac{e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)}}{(2\pi)^{\frac{D}{2}}\sqrt{\det\boldsymbol{\Sigma}}}\pi_c = \dfrac{e^{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}}{(2\pi)^{\frac{D}{2}}\sqrt{\det\boldsymbol{\Sigma}}}\exp\left[-\dfrac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c + \boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \log\pi_c\right]$
  - Let $\boldsymbol{\beta}_c^T = \boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\gamma}_c = \log\pi_c - \dfrac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c$
  - The exponential can then be written as $p(y = c|\boldsymbol{x}) = \dfrac{\exp(\boldsymbol{\beta}_c^T\boldsymbol{x} + \boldsymbol{\gamma}_c)}{\sum_{c'}\exp(\boldsymbol{\beta}_{c'}^T\boldsymbol{x} + \boldsymbol{\gamma}_{c'})}$
    - This is a softmax function
    - The exponential in the softmax makes it so that the largest term dominates while all other terms are usually much smaller
    - Each class has an associated $\boldsymbol{\beta}_c$ and $\boldsymbol{\gamma}_c$, which contains all the info of the class
  - With this, we have $p(y = c|x) \approx \begin{cases} 1 & \boldsymbol{\beta}_c^T\boldsymbol{x} + \boldsymbol{\gamma}_c \gg \boldsymbol{\beta}_{c'}^T\boldsymbol{x} + \boldsymbol{\gamma}_{c'} \\ 0 & \text{otherwise} \end{cases}$
  - The decision rule is $\hat{y}(\boldsymbol{x}) = \operatorname*{argmax}_c \boldsymbol{\beta}_c^T + \boldsymbol{\gamma}_c$
  - This is referred to as *linear Gaussian discriminant analysis*, since the decision boundary is a linear function of $\boldsymbol{x}$
    - The boundary occurs where $\boldsymbol{\beta}_1^T\boldsymbol{x} + \boldsymbol{\gamma}_1 = \boldsymbol{\beta}_0^T\boldsymbol{x} + \boldsymbol{\gamma}_0$ which forms a hyperplane
- More generally, the covariances of the classes are different
  - $\log p(y = c|\boldsymbol{x}) = -\dfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_c) - \dfrac{1}{2}\det\boldsymbol{\Sigma}_c + \log\pi_c - \dfrac{D}{2}\log 2\pi$
  - Consider the boundary between two regions:
    - $-\dfrac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{x} + \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{x} - \dfrac{1}{2}\log\det\boldsymbol{\Sigma}_1 + \log\pi_1 = -\dfrac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{x} + \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{x} - \dfrac{1}{2}\log\det\boldsymbol{\Sigma}_0 + \log\pi_0$
    - $\boldsymbol{x}^T(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_0^{-1})\boldsymbol{x} + 2(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1)^T\boldsymbol{x} + (\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1) + \log\dfrac{\det\boldsymbol{\Sigma}_0}{\det\boldsymbol{\Sigma}_1} + 2\log\dfrac{\pi_1}{\pi_0} = 0$
    - This is a quadratic form that can define a parabola, hyperbola, or even circles and ellipses
- To obtain the parameters of the Gaussian distribution of each class, we can use ML estimation
  - $\hat{\pi}_c = \dfrac{n_c}{n}$ is given by the relative frequency of class c
  - $\hat{\boldsymbol{\mu}}_c = \dfrac{1}{n_c}\sum_i \boldsymbol{x}_i^c$ is given by the sample mean
  - $\hat{\boldsymbol{\Sigma}}_c = \dfrac{1}{n_c}(\boldsymbol{x}_i^c - \hat{\boldsymbol{\mu}}_c)^T(\boldsymbol{x}_i^c - \hat{\boldsymbol{\mu}}_c)$
  - Note that the variance estimate here is biased; use the version with $n_c - 1$ in the denominator for unbiased

Gaussian Parameter Estimation

- Let $\boldsymbol{X}, \boldsymbol{Y}$ be jointly Gaussian and let $\boldsymbol{w} = \begin{bmatrix}\boldsymbol{X} \\ \boldsymbol{Y}\end{bmatrix}$; we want to find the MAP estimator for $\boldsymbol{X}$ given $\boldsymbol{Y}$
  - i.e. we want to find the distribution of $\boldsymbol{X}$ conditioned on $\boldsymbol{Y}$
  - We will make use of the covariances between elements of $\boldsymbol{X}$ and $\boldsymbol{Y}$
  - Strategy: expand out the exponent of the joint PDF and rearrange it into a form with $\boldsymbol{X}$ as the variable

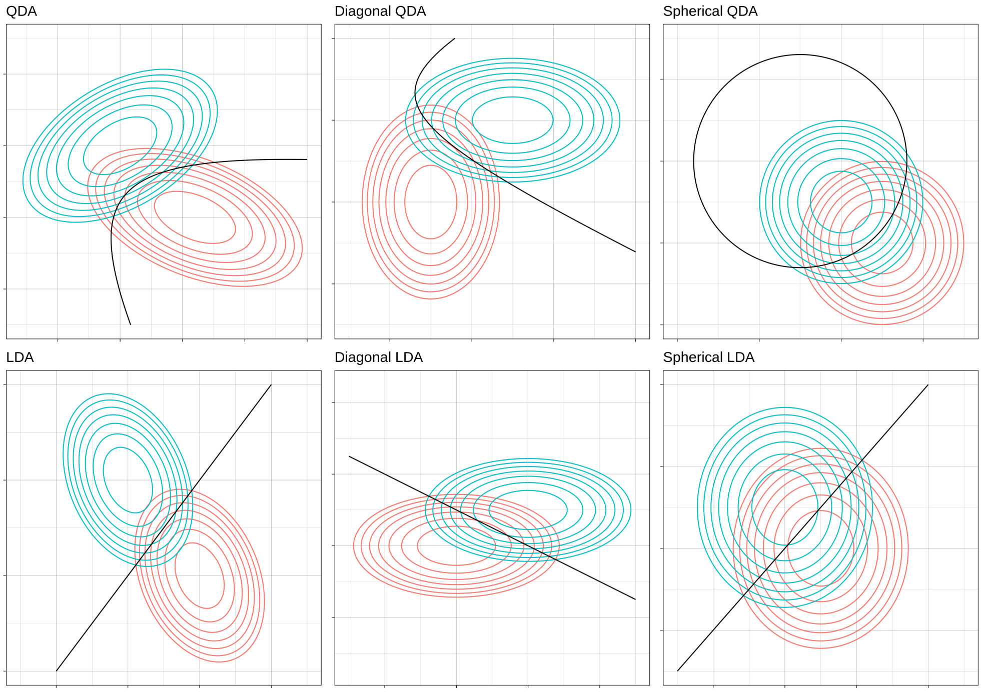

Figure 1: Illustration of the different kinds of Gaussian discriminant analysis.

- The mean of $\boldsymbol{w}$ is $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{bmatrix}$
- The covariance is $\boldsymbol{\Sigma}_w = E[(\boldsymbol{w} - \boldsymbol{\mu}_w)(\boldsymbol{w} - \boldsymbol{\mu}_w)^T] = E\left[\begin{bmatrix} \boldsymbol{X} - \boldsymbol{\mu}_X \\ \boldsymbol{Y} - \boldsymbol{\mu}_Y \end{bmatrix} \begin{bmatrix} (\boldsymbol{X} - \boldsymbol{\mu}_X)^T \\ (\boldsymbol{Y} - \boldsymbol{\mu}_Y)^T \end{bmatrix}\right] = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{XX}} & \boldsymbol{\Sigma}_{\boldsymbol{XY}} \\ \boldsymbol{\Sigma}_{\boldsymbol{YX}} & \boldsymbol{\Sigma}_{\boldsymbol{YY}} \end{bmatrix}$
  - The overall dimension is $N \times N$; the $\boldsymbol{\Sigma}_{\boldsymbol{XY}}, \boldsymbol{\Sigma}_{\boldsymbol{YX}}$ matrices are in general rectangular
  - Let the *precision matrix* $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_w^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{\boldsymbol{XX}} & \boldsymbol{\Lambda}_{\boldsymbol{XY}} \\ \boldsymbol{\Lambda}_{\boldsymbol{YX}} & \boldsymbol{\Lambda}_{\boldsymbol{YY}} \end{bmatrix}$
    * This is the opposite of variance; the larger the precision, the more tightly concentrated the distribution
    * Note $\boldsymbol{\Sigma}_{\boldsymbol{XX}} \neq \boldsymbol{\Lambda}_{\boldsymbol{XX}}^{-1}$, and $\boldsymbol{\Lambda}_{\boldsymbol{XY}} = \boldsymbol{\Lambda}_{\boldsymbol{YX}}^T$
- Now consider the exponent of the joint PDF
  - $$\begin{aligned} &-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_w)^T \boldsymbol{\Sigma}_w^{-1} (\boldsymbol{w} - \boldsymbol{\mu}_w) \\ &= -\frac{1}{2} \begin{bmatrix} \boldsymbol{X} - \boldsymbol{\mu}_X \\ \boldsymbol{Y} - \boldsymbol{\mu}_Y \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Lambda}_{\boldsymbol{XX}} & \boldsymbol{\Lambda}_{\boldsymbol{XY}} \\ \boldsymbol{\Lambda}_{\boldsymbol{YX}} & \boldsymbol{\Lambda}_{\boldsymbol{YY}} \end{bmatrix} \begin{bmatrix} \boldsymbol{X} - \boldsymbol{\mu}_X \\ \boldsymbol{Y} - \boldsymbol{\mu}_Y \end{bmatrix} \\ &= -\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_X)^T \boldsymbol{\Lambda}_{\boldsymbol{XX}} (\boldsymbol{X} - \boldsymbol{\mu}_X) - \frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_X)^T \boldsymbol{\Lambda}_{\boldsymbol{XY}} (\boldsymbol{Y} - \boldsymbol{\mu}_Y) - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{\mu}_Y)^T \boldsymbol{\Lambda}_{\boldsymbol{YX}} (\boldsymbol{X} - \boldsymbol{\mu}_X) - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{\mu}_Y)^T \\ &= -\frac{1}{2}\left(\boldsymbol{X}^T \boldsymbol{\Lambda}_{\boldsymbol{XX}} \boldsymbol{X} - \boldsymbol{X}^T \boldsymbol{\Lambda}_{\boldsymbol{XY}} (\boldsymbol{Y} - \boldsymbol{\mu}_Y) + (\boldsymbol{Y} - \boldsymbol{\mu}_Y)^T \boldsymbol{\Lambda}_{\boldsymbol{YX}} \boldsymbol{X} - \boldsymbol{\mu}_X^T \boldsymbol{\Lambda}_{\boldsymbol{XX}} \boldsymbol{X} - \boldsymbol{X}^T \boldsymbol{\Lambda}_{\boldsymbol{XX}} \boldsymbol{\mu}_X + \dots\right) \\ &= -\frac{1}{2}\left(\boldsymbol{X}^T \boldsymbol{\Lambda}_{\boldsymbol{XX}} \boldsymbol{X} - 2\boldsymbol{X}^T (\boldsymbol{\Lambda}_{\boldsymbol{XX}} \boldsymbol{\mu}_X - \boldsymbol{\Lambda}_{\boldsymbol{XY}} (\boldsymbol{Y} - \boldsymbol{\mu}_Y)) + \dots\right) \end{aligned}$$
  - This gives us $f(\boldsymbol{x}|\boldsymbol{y})$ and implies that it is jointly Gaussian
  - Let $f(\boldsymbol{x}|\boldsymbol{y}) = c \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{X|Y})^T \boldsymbol{\Sigma}_{X|Y}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{X|Y})\right] = c \exp\left[-\frac{1}{2}\left(\boldsymbol{x} \sigma_{\boldsymbol{XY}}^{-1} \boldsymbol{x} - 2\boldsymbol{x}^T \boldsymbol{\Sigma}_{\boldsymbol{XY}}^{-1} \boldsymbol{\mu}_{X|Y}\right)\right]$
  - By matching terms we see $\boldsymbol{\Sigma}_{\boldsymbol{XY}} = \boldsymbol{\Lambda}_{\boldsymbol{XX}}^{-1}$ and $\boldsymbol{\mu}_{X|Y} = \boldsymbol{\mu}_X - \boldsymbol{\Lambda}_{\boldsymbol{XX}}^{-1} \boldsymbol{\Lambda}_{\boldsymbol{XY}} (\boldsymbol{y} - \boldsymbol{\mu}_Y)$
  - Therefore $f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\mu}_{X|Y}, \boldsymbol{\Sigma}_{X|Y})$
- Given this PDF, we can see that the MAP estimate is simply $\boldsymbol{\mu}_{X|Y}$
  - We can show that this is the same as the LMS estimate

- However, we only have this in terms of the precision matrix; can we find it in terms of $\boldsymbol{\Sigma}$?

- Using the Schur complement on $\boldsymbol{\Sigma}^{-1}$ we can find a general expression for each of the $\boldsymbol{\Lambda}$
  - $\boldsymbol{\Lambda}_{\boldsymbol{XX}} = (\boldsymbol{\Sigma}_{\boldsymbol{XX}} - \boldsymbol{\Sigma}_{\boldsymbol{XY}}\boldsymbol{\Sigma}_{\boldsymbol{YY}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{YX}})^{-1}$
  - $\boldsymbol{\Lambda}_{\boldsymbol{XY}} = (\boldsymbol{\Sigma}_{\boldsymbol{XX}} - \boldsymbol{\Sigma}_{\boldsymbol{XY}}\boldsymbol{\Sigma}_{\boldsymbol{YY}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{YX}})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{XY}}\boldsymbol{\Sigma}_{\boldsymbol{YY}}^{-1}$
  - Therefore $\boldsymbol{\mu}_{X|Y} = \boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{\boldsymbol{XY}}\boldsymbol{\Sigma}_{\boldsymbol{YY}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_Y)$, $\boldsymbol{\Sigma}_{X|Y} = \boldsymbol{\Sigma}_{\boldsymbol{XX}} - \boldsymbol{\Sigma}_{\boldsymbol{XY}}\boldsymbol{\Sigma}_{\boldsymbol{YY}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{YX}}$
  - We get both the mean of the estimate and its sprea