UNIVERSITY OF CALIFORNIA, BERKELEY

Homework 3: CS 189

Oscar Ortega

July 16, 2021

1 GAUSSIAN CLASSIFICATION

a:

The Bayes optimal decision boundary is found when the probabilities $P(x|C_2) = P(x|C_1)$ as it is the case $P(C_2) = P(C_1)$

WLOG we allow $\mu_2 > \mu_1$

$$P(x|C_2) = P(x|C_1)$$

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

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

$$x^2 - 2x\mu_2 + \mu_2^2 = x^2 - 2x\mu_1 + \mu_1^2$$

$$x(\mu_2 - \mu_1) = \frac{1}{2}\mu_2^2 - \mu_1^2$$

$$x = \frac{\mu_2^2 - \mu_1^2}{2(\mu_2 - \mu_1)}$$

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

$$r^*_{\text{gauss}}(x) = \{C_1 \text{ if } P(x|C_1) \le \frac{\mu_2 + \mu_1}{2} : C_2 \text{ otherwise }\}$$

b: Bayes Error is $P_e = P(\text{ misclassified as } C_1|C_2)P(C_2) + P(\text{ misclassified as } C_2|C_1)P(C_1)$

$$\frac{1}{2}\left(P(x \leq \frac{\mu_2+\mu_1}{2}|C_2) + P(x \geq \frac{\mu_2+\mu_1}{2}|C_1)\right)$$

$$\frac{1}{2}\left(P(x-\mu_2 \leq \frac{\mu_2+\mu_1}{2} - \mu_2) + P(x-\mu_1 \geq \frac{\mu_2+\mu_1}{2} - \mu_1)\right)$$

$$\frac{1}{2}\left(P(\frac{x-\mu_2}{\sigma} \leq \frac{-\mu_2+\mu_1}{2\sigma}) + P(\frac{x-\mu_1}{\sigma} \geq \frac{\mu_2-\mu_1}{2\sigma})\right)$$

$$\frac{1}{2}\left(\Phi(\frac{-\mu_2+\mu_1}{2\sigma}) + Q(\frac{\mu_2-\mu_1}{2\sigma})\right)$$

$$Q(\frac{\mu_2-\mu_1}{2\sigma}) = \int_a^\infty e^{\frac{-z^2}{2}}dz$$

Where $a = \frac{\mu_2-\mu_1}{2\sigma}$

2 Isocontours of Normal Distributions

This portion is on the IPython Notebook.

3 Eigenvectors of the Gaussian Covariance Matrix

This portion is on the IPython Notebook.

4 Classification

for $r : \mathbf{R}^d \to \{1,...,c+1\}$ Let $\mathscr{L}(r(x) = i, y = j) = \{0$ if $i = j : \lambda_r$ if i = c + 1 : $\lambda_s$otherwise $\}$
We define the risk as follows : (a):

$$R(r(x) = i|x) = \mathbf{E}(L(r(x), Y) = \sum_{j=1}^{c} L(r(x) = i, y = j)P(Y = j|x)$$

WLOG $\lambda_r \leq \lambda_s$ and let $\lambda_s > 0$:
If we choose class i

$$R(r(x) = i \in \{1,...,c\}|x) = \sum_{j=1}^{c} L(r(x) = i, y = j)P(Y = j|x)$$

$$= \sum_{j=1}^{c} \lambda_s P(Y = j|x) \leftarrow \text{ because we are only encountering loss from } \lambda_s$$

$$\lambda_s(1 - P(Y = i|x))$$

If we choose doubt

$$R(r(x) = c+1|x) = \sum_{i=1}^{c} \lambda_r P(y = i|x)$$

$$= \lambda_r$$

So if we want to optimize the decision rule when choosing a class i we want to insure that choosing i is better than choosing either doubt or another class.
Case 1: choosing i over another class j

$$R(r(x) = i|x) \leq R(r(x) = j, j \neq i)$$

$$\lambda_s(1 - P(Y = i|x) \leq \lambda_s(1 - P(Y = j|x)$$

$$P(Y = i|x) \geq P(Y = j|x)$$

Similarily, we also want to know to ensure optimality of choosing a class i over choosing doubt.

$$R(r(x) = i|x) \leq R(r(x) = c + 1|x)$$

$$\lambda_s(1 - P(Y = i|x) \leq \lambda_r$$

$$P(Y = i|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$$

Therefore, we know we can choose class i optimally if the two conditions are met $\forall i$.

(b): $\lim \lambda_r = 0 \rightarrow$ we choose class i if $P(Y = i|x) \geq P(Y = j|x)$ for all j and $P(Y = i|x)) \geq 1$, which is only true in the case we are one hundred percent certain about classifying the data point. Similarly if we allow $\lambda_r \geq \lambda_s \rightarrow$ we choose class i if $P(Y = i|x) \geq P(Y = j|x)$ for all j and if $P(Y = i|x) \geq 0$. Intuitively, this is what should be happening because in the case you no longer punish by setting λ_r equal to 0, it becomes more and more unlikely for the policy to decide on ever guessing a class. Similarly, If we allow $\lambda_r \geq \lambda_s$, this corresponds to making it less desirable to remain in doubt than to make a guess which is why we only need to make sure the class is the class that occurred with maximum probability.

5 Maximum Likelihood Estimation

a: Let $Y := \mathcal{N}(\mu, \Sigma)$

$$Pr(X_1, X_2, ...X_N|Y) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\right)$$

Because ln is a monotonically increasing function, maximizing the probability of the samples is the same as maximizing the natural log of the probability.

$$\ln(Pr(X_1, X_2, ...X_N|Y)) = \ln\left(\prod_{i=1}^{N} \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\right)\right)$$

$$= N \ln\left(\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\right) - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

We can drop the terms involving π because they do not affect our maximization. So now our goal is to minimize the following.

$$Q(\Sigma,\mu) = -\frac{N}{2}ln(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

$$\frac{\partial Q}{\partial \mu} = \sum_{i=1}^{n}\Sigma^{-1}(x_i-\mu) = 0$$

$$n\Sigma^{-1}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\sum_{i=1}^{n}x_i = n\mu$$

$$\frac{\sum_{i=1}^{n}x_i}{n} = \hat{\mu}$$

$$Q(\Sigma,\mu) = -\frac{N}{2}ln(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

$$Q(\Sigma,\mu) = -\frac{N}{2}ln(\Pi_{i=1}^{n}\sigma_i^2) - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

$$Q(\Sigma,\mu) = -\frac{N}{2}\sum_{i=1}^{N}ln(\sigma_i^2) - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

$$= -N^2\sum_{i=1}^{N}\ln(\sigma_i) - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

$$\frac{\partial Q}{\partial \sigma_k} = \frac{-N^2}{\sigma_k} + \frac{N}{\sigma_k^3}\sum_{i=1}^{N}(x_i-\mu)^T(x_i-\mu) = 0$$

$$\frac{N^2}{\sigma_k} = \frac{N}{\sigma_k^3}\sum_{i=1}^{N}(x_i-\mu)^2$$

$$\hat{\sigma_k}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

b:

$$Q(\Sigma,\mu) = -\frac{N}{2}ln(\Pi_{i=1}^{n}\sigma_i^2) - \frac{1}{2}\sum_{i=1}^{N}(x_i-A\mu)^T\Sigma^{-1}(x_i-A\mu)$$

$$\frac{\partial Q}{\partial \mu} = -\frac{1}{2}\sum_{i=1}^{N} 2\Sigma^{-1}(x_i - A\mu)\frac{\partial(x_i - A\mu)}{\partial \mu}$$

$$\frac{1}{2}\sum_{i=1}^{N} 2\Sigma^{-1}(Ix_i - A\mu)A = 0$$

$$N\Sigma^{-1}A\sum_{i=1}^{N} x_i + N\Sigma^{-1}A\sum_{i=1}^{N} -Am = 0$$

$$N\Sigma^{-1}A\sum_{i=1}^{N} x_i - N^2\Sigma^{-1}A^2\mu = 0$$

$$\frac{A^{-1}\sum_{i=1}^{N} x_i}{N} = \hat{\mu}$$

6 Covariance Matrices and Decomposition

(A) $\hat{\Sigma}$ is singular if certain features of the data, are deterministically a function of other features of the data, This corresponds to when a feature x_i has no variance, i.e the diagonal entries of the covariance matrix are 0 for at least one feature. Geometrically, if we consider let $\Sigma \in \mathbb{R}^{n,n}$ with $dim(Range(\Sigma)) = r : r < n$ we can consider the features that are in the range of the covariance matrix as forming the support of an r dimensional gaussian, and the remaining $n - r$ features as affine functions of the gaussian.

(B) If we have a singular covariance matrix estimator, $\hat{\Sigma}$, one way we can make this matrix invertible is by multiplying $\hat{\Sigma}$ by λI where λ is small. λ, like any other hyperparameter can be optimized through the performance on validation sets.
(C) If we recall that the matrix $\hat{\Sigma}$ is symettric, this implies the following:

$$\hat{\Sigma} = U\Lambda U^T$$

Where U is an orthonormal basis of eigenvectors corresponding to the eigenvalues in the diagonal matrix Λ, ie $\Sigma u_i = \lambda_i u_i$. Because it is the case that U^T forms a basis for the column space of Σ, we can represent x as a linear combination of the U vectors. Furthermore,

$$\Sigma^{-1} = U\Lambda^{-1}U^T$$

So if we want to maximize the pdf function which is equivalent to maximizing $x^T\Sigma^{-1}x$.

$$= argmax_x x^T U\Lambda U^T x$$

Defining $x' = U^T x$

$$= argmax_{x'} x'^T \Lambda^{-1} x'$$

$$= u_n$$

the eigenvector corresponding to the smallest eigenvalue of Λ, the largest eigenvalue of Λ^{-1}