Exercício 1: Análise da Fronteira de Decisão em Modelos Generativos

(a) Fronteira de Decisão com Covariâncias Diferentes e Priors Iguais

A probabilidade posterior de uma classe C_k dado um vetor de features $\mathbf{x}$ é dada pelo Teorema de Bayes:

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})}.$$

A fronteira de decisão ocorre quando $p(C_1|\mathbf{x}) = p(C_2|\mathbf{x})$. Assumindo priors iguais $p(C_1) = p(C_2) = 0.5$, esta condição se simplifica para $p(\mathbf{x}|C_1) = p(\mathbf{x}|C_2)$.

As densidades condicionais de classe são Gaussianas Multivariadas:

$$p(\mathbf{x}|C_k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right).$$

Igualando $p(\mathbf{x}|C_1) = p(\mathbf{x}|C_2)$ e tomando o logaritmo natural de ambos os lados, obtemos:

$$\begin{aligned}
&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_1| \\
&= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_2|.
\end{aligned}$$

Multiplicando por -2 e rearranjando, a equação da fronteira de decisão é:

$$(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \ln|\boldsymbol{\Sigma}_1| = (\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) + \ln|\boldsymbol{\Sigma}_2|.$$

Expandindo os termos quadráticos, temos:

$$\mathbf{x}^T\boldsymbol{\Sigma}_1^{-1}\mathbf{x} - 2\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\mathbf{x} + \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \ln|\boldsymbol{\Sigma}_1| = \mathbf{x}^T\boldsymbol{\Sigma}_2^{-1}\mathbf{x} - 2\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1}\mathbf{x} + \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 + \ln|\boldsymbol{\Sigma}_2|.$$

Reorganizando os termos envolvendo $\mathbf{x}$, obtemos:

$$\mathbf{x}^T(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} - 2(\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 + \ln|\boldsymbol{\Sigma}_1| - \ln|\boldsymbol{\Sigma}_2|) = 0.$$

Esta é uma função quadrática de $\mathbf{x}$, e a forma geométrica desta fronteira é uma **hiperquadrica**. Casos específicos incluem hiperboloides, paraboloides ou elipsoides, dependendo das matrizes de covariância e dos vetores de média.

(b) Efeito da Alteração do Prior com Covariâncias Compartilhadas

Se as covariâncias são compartilhadas, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, e os priors são $p(C_1) = \pi$ e $p(C_2) = 1-\pi$, a condição para a fronteira de decisão $p(\mathbf{x}|C_1)p(C_1) = p(\mathbf{x}|C_2)p(C_2)$ torna-se:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\pi = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})(1-\pi).$$

Tomando o logaritmo natural de ambos os lados:

$$\begin{aligned}
&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \ln\pi \\
&= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \ln(1-\pi).
\end{aligned}$$

Os termos envolvendo $|\mathbf{\Sigma}|$ se cancelam, e expandindo os termos quadráticos:

$$-\frac{1}{2}(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x} - 2\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1) + \ln\pi = -\frac{1}{2}(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x} - 2\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2) + \ln(1-\pi).$$

Os termos $\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}$ também se cancelam. Multiplicando por -2 e rearranjando, obtemos a fronteira de decisão linear:

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2) + \ln\left(\frac{\pi}{1-\pi}\right) = 0.$$

O termo $\ln\left(\frac{\pi}{1-\pi}\right)$ é um deslocamento escalar na equação linear. Geometricamente, **alterar o prior π resulta em um deslocamento paralelo da fronteira de decisão linear**.

- Se $\pi > 0.5$ (ou seja, $p(C_1) > p(C_2)$), então $\ln\left(\frac{\pi}{1-\pi}\right) > 0$. Para satisfazer a equação, o valor de $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}\mathbf{x}$ precisa ser menor para um dado $\mathbf{x}$, o que significa que a fronteira se desloca em direção à região onde $p(\mathbf{x}|C_1)$ é menor em relação a $p(\mathbf{x}|C_2)$. Intuitivamente, como C_1 é mais provável a priori, a região classificada como C_1 se expande.

- Se $\pi < 0.5$ (ou seja, $p(C_1) < p(C_2)$), então $\ln\left(\frac{\pi}{1-\pi}\right) < 0$, e a fronteira se desloca na direção oposta, expandindo a região de C_2.

- Se $\pi = 0.5$, o termo logarítmico é zero, e a fronteira é determinada apenas pelas médias e pela covariância compartilhada.

(c) Uso da Distância de Mahalanobis para Classificação com Dois Modelos Gaussianos

Se tivéssemos um modelo Gaussiano para a classe normal (C_1) com parâmetros $\boldsymbol{\mu}_1, \mathbf{\Sigma}_1$ e outro para a classe fraude (C_2) com parâmetros $\boldsymbol{\mu}_2, \mathbf{\Sigma}_2$, a distância de Mahalanobis $\Delta_k^2 = (\mathbf{x} - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)$ para cada classe poderia ser usada para classificação com base na probabilidade posterior.

Uma abordagem seria usar a distância de Mahalanobis para calcular a densidade de probabilidade para cada classe e então usar o Teorema de Bayes para determinar a probabilidade posterior de cada classe dado um novo ponto $\mathbf{x}$:

$$p(C_k|\mathbf{x}) \propto p(\mathbf{x}|C_k)p(C_k) \propto \frac{1}{|\mathbf{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}\Delta_k^2\right) p(C_k).$$

Poderíamos então classificar $\mathbf{x}$ na classe com a maior probabilidade posterior. Se assumirmos priors iguais, $p(C_1) = p(C_2)$, a classificação se basearia em qual das densidades $p(\mathbf{x}|C_k)$ é maior, o que está relacionado à distância de Mahalanobis e ao determinante da matriz de covariância.

Outra forma seria definir um limiar nas distâncias de Mahalanobis. Por exemplo, se modelarmos a classe normal (C_1) e considerarmos qualquer ponto com uma distância de Mahalanobis acima de um certo limiar como uma anomalia (fraude), isso implicitamente assume algo sobre a distribuição da classe fraude. Se tivéssemos um modelo para ambas as classes, poderíamos comparar suas distâncias relativas, possivelmente ponderadas pelos priors.

As premissas implícitas ao usar modelos Gaussianos e a distância de Mahalanobis para classificação neste cenário incluem:

- **Cada classe (normal e fraude) pode ser adequadamente modelada por uma distribuição Gaussiana multivariada.** Isso significa que os dados dentro de cada classe estão concentrados em torno de uma média e a dispersão pode ser descrita por uma matriz de covariância.

- **As matrizes de covariância $\mathbf{\Sigma}_1$ e $\mathbf{\Sigma}_2$ são inversíveis (não singulares).** Isso é necessário para calcular a distância de Mahalanobis.

- **Se priors diferentes forem usados, eles devem ser estimados ou conhecidos.** A decisão ótima depende não apenas das probabilidades condicionais, mas também das probabilidades a priori das classes.

- **A fronteira de decisão baseada em modelos Gaussianos (como derivado em (a)) é apropriada para separar as classes.** Se as verdadeiras distribuições das classes se desviarem significativamente da forma Gaussiana, essa fronteira pode não ser ótima.

Em essência, ao usar a distância de Mahalanobis dentro de um arcabouço Bayesiano com modelos Gaussianos, estamos assumindo que a probabilidade de um ponto pertencer a uma classe diminui exponencialmente com o quadrado da sua distância de Mahalanobis do centro daquela classe, levando em conta a forma da distribuição (através da matriz de covariância).