PRÁCTICA 6: Recoñecemento xeométrico de formas 2D con clasificadores Bayesianos

Xosé R. Fdez-Vidal
xose.vidal@usc.es

Grao de Robótica. EPSE de Lugo — 8 de decembro do 2022

1. Introdución

O recoñecemento de patróns é a identificación automática de de regularidades nos datos. Ten aplicacións en diversos campos dende a análise estatística, procesamento de sinais, análise de imaxes, recuperación de información, bioinformática, compresión de datos, gráficos por ordenador e aprendizaxe automática. O recoñecemento de patróns ten a súa orixe na estatística e na enxeñaría; Algúns enfoques modernos para o recoñecemento de patróns inclúen o uso da aprendizaxe automática, debido á maior dispoñibilidade de big data e maior capacidade de procesamento. Estas actividades pódense ver como dúas facetas do mesmo campo de aplicación, e experimentaron un desenvolvemento substancial nas últimas décadas.

En aprendizaxe automática, o recoñecemento de patróns consisten na asignación dunha etiqueta a un valor de entrada determinado. Na estatística, a análise discriminante introduciuse para este mesmo propósito en 1936. Un exemplo de recoñecemento de patróns é a *clasificación*, que tenta asignar cada valor de entrada a un determinado conxunto de clases.

Os algoritmos de recoñecemento de patróns, xeralmente, pretenden proporcionar unha resposta razoable para todas as entradas posibles e realizar a correspondencia "máis probable" das entradas, tendo en conta a súa variación estatística. Isto oponse aos algoritmos de "matching", que buscan coincidencias exactas entre a entrada e os patróns preexistentes.

Unha definición moderna de recoñecemento de patróns é:

O campo do recoñecemento de patróns ocúpase do descubrimento automático de regularidades nos datos mediante o uso de algoritmos informáticos e co uso destas regularidades para realizar accións como a clasificación dos datos en diferentes categorías.

O recoñecemento de patróns clasifícase xeralmente segundo o tipo de procedemento de aprendizaxe utilizado para xerar o valor de saída. A aprendizaxe supervisada supón que se proporcionou un conxunto de datos de adestramento, consistente nun conxunto de instancias que foron etiquetadas correctamente a man coa saída correcta. Un procedemento de aprendizaxe xera un modelo que tenta cumprir dous obxectivos ás veces conflitivos: obter bos resultados cos datos de adestramento e ser capaz de xeneralizar, o mellor posible, con novos datos cos que nunca tivo contacto (fase test). A aprendizaxe sen supervisión, por outra banda, asume datos de adestramento sen ser etiquetados a man, e tenta atopar patróns inherentes nos datos que logo se poden usar para determinar o valor de saída correcto para novas instancias. Unha combinación dos dous anteriores enfoques é a aprendizaxe semisupervisada, que utiliza unha combinación de datos etiquetados e sen etiquetar (normalmente un pequeno conxunto de datos etiquetados combinados cunha gran cantidade de datos sen etiquetar).

Ás veces utilízanse termos diferentes para describir os correspondentes procedementos de aprendizaxe supervisada e non supervisada para o mesmo tipo de saída. O equivalente non supervisado da clasificación coñécese normalmente como agrupamento (clustering), que tenta agrupar os datos de entrada en conxuntos (clústeres) en función dalgunha medida de semellanza inherente (por exemplo, a distancia entre instancias, considerada como vectores nun espazo vectorial multidimensional), en lugar de asignar cada instancia de entrada a unha clase pertencnete a un determinado conxunto predefinido.

O dato de entrada para o que se xera un valor de saída denomínase formalmente *instancia* e descríbese formalmente por un *vector de características*, que en conxunto constitúen unha descrición de todas os aspectos coñecidos da instancia. Estes vectores de características pódense ver como puntos que definen un espazo multidimensional, e pódense aplicar correspondentes métodos para manipulalos como vectores

en espazos vectoriais. As características adoitan ser categóricas (tamén coñecidas como nominais, é dicir, que consisten nun conxunto de elementos non ordenados, como un xénero d "masculin" ou "feminino", ou un tipo sanguíneo de "A", "B", "AB" ou "0"), ordinal (que consiste nun conxunto de elementos ordenados, por exemplo, "grande","mediano" ou "pequeno"), con valor enteiro (por exemplo, un reconto do número de ocorrencias dunha palabra concreta nun documento) ou de valor real (por exemplo, unha medida da presión arterial). Moitas veces, os datos categóricos e ordinais agrúpanse, e este tamén é o caso dos datos con valores enteiros e con valores reais. Moitos algoritmos só funcionan en termos de datos categóricos e requiren que os datos con valores reais ou enteiros sexan discretizados en grupos (por exemplo, menos de 5, entre 5 e 10 ou maiores de 10).

2. Formulación do problema

O problema do recoñecemento de patróns pódese enunciar do seguinte xeito: Dada unha función descoñecida $g : \mathcal{X} \to \mathcal{Y}$ (o ground truth) que mapea instancias de entrada $\boldsymbol{x}\ en\mathcal{X}$ para emitir as etiquetas $y \in \mathcal{Y}$, xunto cos datos de adestramento $\mathbf{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ que se supón que representan exemplos precisos da asignación, produce unha función $h : \mathcal{X} \to \mathcal{Y}$ que se aproxime o máis posible á correspondencia correcta g. Por exemplo, se o problema é filtrar o correo lixo, entón $\boldsymbol{x}_i$ é algunha representación dun correo electrónico e y é"spam" ou "non spam"). Para que este sexa un problema ben definido, é preciso definir con rigor "as aproximacións o máis posible". Na teoría da decisión, isto defínese especificando unha función de perda ou función de custo que asigna un valor específico á "perda" resultante de producir unha etiqueta incorrecta. O obxectivo entón é minimizar a perda esperada, tomando a expectativa sobre a distribución de probabilidade de $\mathcal{X}$. Na práctica, nin a distribución de $\mathcal{X}$ nin a función de ground truth $g : \mathcal{X} \to \mathcal{Y}$ coñécense con exactitude, pero si se poden calcular empiricamente recollendo un gran número de mostras de $\mathcal{X}$ e etiquetándoas manualmente usando o valor correcto de $\mathcal{Y}$[1]. A función de perda particular depende do tipo de etiqueta que se prevé. Por exemplo, no caso da clasificación, a función simple de perdas cero-un é a miúdo suficiente. Isto corresponde simplemente a asignar unha perda de 1 a calquera etiquetaxe incorrecta e implica que o clasificador óptimo minimiza a taxa de erro nos datos de proba independentes (é dicir, contando a fracción de instancias que a función aprendida) $h : \mathcal{X} \to \mathcal{Y}$ etiquetas incorrectamente, o que equivale a maximizar o número de instancias clasificadas correctamente). O obxectivo do procedemento de aprendizaxe é entón minimizar a taxa de erro (maximizar a corrección) nun conxunto de probas "típicos".

Para un recoñecedor de patróns probabilísticos, o problema é estimar a probabilidade de cada posible etiqueta de saída dada unha instancia de entrada particular, é dicir, estimar unha función da forma

$$p(\text{label}|\boldsymbol{x}, \boldsymbol{\theta}) = f\left(\boldsymbol{x}; \boldsymbol{\theta}\right). \tag{1}$$

onde a entrada do vector de características é $\boldsymbol{x}$, e a función f normalmente está parametrizada por algúns parámetros $\boldsymbol{\theta}$. Nun enfoque discriminativo do problema, f estímase directamente. Nunha aproximación xerativa, con todo, a probabilidade inversa $p(\boldsymbol{x}|\text{label})$ estímase e combínase coa probabilidade previa $p(\text{label}|\boldsymbol{\theta})$ usando a regra de Bayes, como segue:

$$p(\text{label}|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|\text{label}, \boldsymbol{\theta})p(\text{label}|\boldsymbol{\theta})}{\sum_{L \in \text{all labels}} p(\boldsymbol{x}|L)p(L|\boldsymbol{\theta})}. \tag{2}$$

Cando as etiquetas se distribúen continuamente (por exemplo, na análise de regresión), o denominador implica a integración máis que a suma:

$$p(\text{label}|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|\text{label}, \boldsymbol{\theta})p(\text{label}|\boldsymbol{\theta})}{\int_{L \in \text{all labels}} p(\boldsymbol{x}|L)p(L|\boldsymbol{\theta})\,\mathrm{d}\,L}. \tag{3}$$

O valor de $\boldsymbol{\theta}$ apréndese normalmente mediante a estimación máxima a posteriori (MAP). Deste xeito atopa o mellor valor que satisface simultaneamente dous obxectos en conflito: para realizar o mellor posible nos datos de adestramento (menor taxa de erro) e para atopar o modelo máis sinxelo posible. Esencialmente, isto combina a estimación de máxima verosimilitude cun procedemento de regularización que favorece modelos máis sinxelos fronte a modelos máis complexos. Nun contexto bayesiano, o procedemento de regularización pódese ver como colocar unha probabilidade $p(\boldsymbol{\theta})$ en diferentes valores de $\boldsymbol{\theta}$. Matematicamente:

[1] un proceso lento, que adoita ser o factor limitante na cantidade de datos deste tipo que se poden recoller

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{D}) \tag{4}$$

onde $\boldsymbol{\theta}^*$ é o valor usado para $\boldsymbol{\theta}$ no procedemento de avaliación posterior, e $p(\boldsymbol{\theta}|\mathbf{D})$, a probabilidade posterior de $\boldsymbol{\theta}$, vén dada por

$$p(\boldsymbol{\theta}|\mathbf{D}) = \left[\prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})\right] p(\boldsymbol{\theta}). \tag{5}$$

No enfoque bayesiano deste problema, en lugar de escoller un único vector de parámetros, a probabilidade de a etiqueta dada para unha nova instancia calcúlase $x\boldsymbol{x}$ integrando todos os valores posibles de $\boldsymbol{\theta}$, ponderado segundo a probabilidade posterior:

$$p(\text{label}|\boldsymbol{x}) = \int p(\text{label}|\boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{D}) \, \mathrm{d}\, \boldsymbol{\theta}. \tag{6}$$

3. Clasificador bayesiano e casos particulares

Como dixemos anteriormente, a regra de clasificación máis empregada é aquela que minimiza o erro de toda a clasificación. A regra que cumpre con esta premisa é a regra de Bayes, permite asignar ao obxecto a clase que teña a maior probabilidade. Formalmente, se temos un conxunto de grupos a regra de Bayes asigna o obxecto á clase i-ésima tal que:

$$P(G_i|\vec{x}) > P(G_j|\vec{x}) \; \forall i \neq j. \tag{7}$$

O inconveniente é que as anteriores probabilidades condicionais son descoñecidas e non existen métodos estándar para estimalas. Sen embargo, é posible estimar a probabilidade condicional da clase G_i o vector de características $\vec{x}$. Esta probabilidade $P(\vec{x}|G_i)$ estímase collendo un número de mostras da clase i-ésima y calculándoa para clase i.

O teorema de Bayes establece a conexión entre estes dous topos de probabilidades condicionais:

$$P(G_i|\vec{x}) = \frac{P(\vec{x}|G_i)P(G_i)}{\sum_k^{N_{clases}} P(\vec{x}|G_k)P(G_k)} \tag{8}$$

onde $P(G_i)$ é a probabilidade do que o obxecto pertenza á clase G_i, é dicir a proporción da clase G_i na poboación total. Empregando o teorema de Bayes na regra de clasificación obtemos:

$$\frac{P(\vec{x}|G_i)P(G_i)}{\sum_k P(\vec{x}|G_k)P(G_k)} > \frac{P(\vec{x}|G_j)P(G_j)}{\sum_k P(\vec{x}|G_k)P(G_k)} \forall j \neq j, \tag{9}$$

de onde

$$P(\vec{x}|G_i)P(G_i) > P(\vec{x}|G_j)P(G_j) \forall j \neq j, \tag{10}$$

Os clasificadores estatísticos que se basean na regra de Bayes denomínanse clasificadores de bayesianos e constitúen a aproximación fundamental estatística ao recoñecemento de patróns.

A estrutura dun clasificador bayesiano esta determinada principalmente polas funcións de densidade de probabilidade $P(x|G_i)$. De todas as funcións de densidade de probabilidade estudadas, a que máis atención recibiu é a normal:

$$P(x|G_i) = \mathcal{N}(\vec{x}; \vec{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi^n \det(\Sigma_y)}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})} \tag{11}$$

Sempre que asumimos unha determinada forma das densidades de probabilidade estamos, dunha maneira implícita, aceptando un modelo válido baixo unhas determinadas condicións.

Observa que neste caso, a probabilidade $P(x|G_i)$ para a distribución normal queda completamente estimada para a clase G_i determinado dous parámetros: $\vec{\mu}_i$ vector media da clase G_i e a matriz Σ_i de covarianza da mesma clase. Desta forma, a regra de Bayes pode expresase como:

$$\frac{P(G_i)}{\sqrt{2\pi^n \det(\Sigma_i)}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu_i})^T {\Sigma_i}^{-1} (\vec{x}-\vec{\mu_i})} > \frac{P(G_j)}{\sqrt{2\pi^n \det(\Sigma_j)}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu_j})^T {\Sigma_j}^{-1} (\vec{x}-\vec{\mu_j})}. \tag{12}$$

Aplicando logaritmos neperianos e eliminando termos comúns:

$$-\ln(\det(\Sigma_i)-0{,}5\cdot(\vec{x}-\vec{\mu_i})^T{\Sigma_i}^{-1}(\vec{x}-\vec{\mu_i})+\ln(P(G_i)) > -\ln(\det(\Sigma_j)-0{,}5\cdot(\vec{x}-\vec{\mu_j})^T{\Sigma_j}^{-1}(\vec{x}-\vec{\mu_j})+\ln(P(G_j)). \tag{13}$$

Se multiplicamos por -1 ámbolos dous membros e cambiamos o signo:

$$\ln(\det(\Sigma_i)+0{,}5\cdot(\vec{x}-\vec{\mu_i})^T{\Sigma_i}^{-1}(\vec{x}-\vec{\mu_i})-\ln(P(G_i)) < \ln(\det(\Sigma_j)-0{,}5\cdot(\vec{x}+\vec{\mu_j})^T{\Sigma_j}^{-1}(\vec{x}-\vec{\mu_j})-\ln(P(G_j)). \tag{14}$$

Nestas condicións, a función discriminante para a clase G_i é:

$$d_i(X) = \ln(\det(\Sigma_i)) + 0{,}5 \cdot (\vec{x}-\vec{\mu_i})^T{\Sigma_i}^{-1}(\vec{x}-\vec{\mu_i}) - \ln(P(G_i)), \tag{15}$$

onde o centro da clase está dado por $\vec{\mu_i}$ e a distribución de cada clase se caracteriza pola matriz de covarianza (Fig. 1). Pode demostrarse que o lugar xeométrico dos puntos con densidade constante son hiperelipsoides para os cales a forma cadrática $r^2 = (\vec{x}-\vec{\mu_i})^T{\Sigma_i}^{-1}(\vec{x}-\vec{\mu_i})$ é constante. A dirección dos eixos principais destes hiperelipsoides están dados polo vectores propios de Σ_i, mentres que as súas lonxitudes de eixos veñen determinadas pola distancia de Mahalanobis.

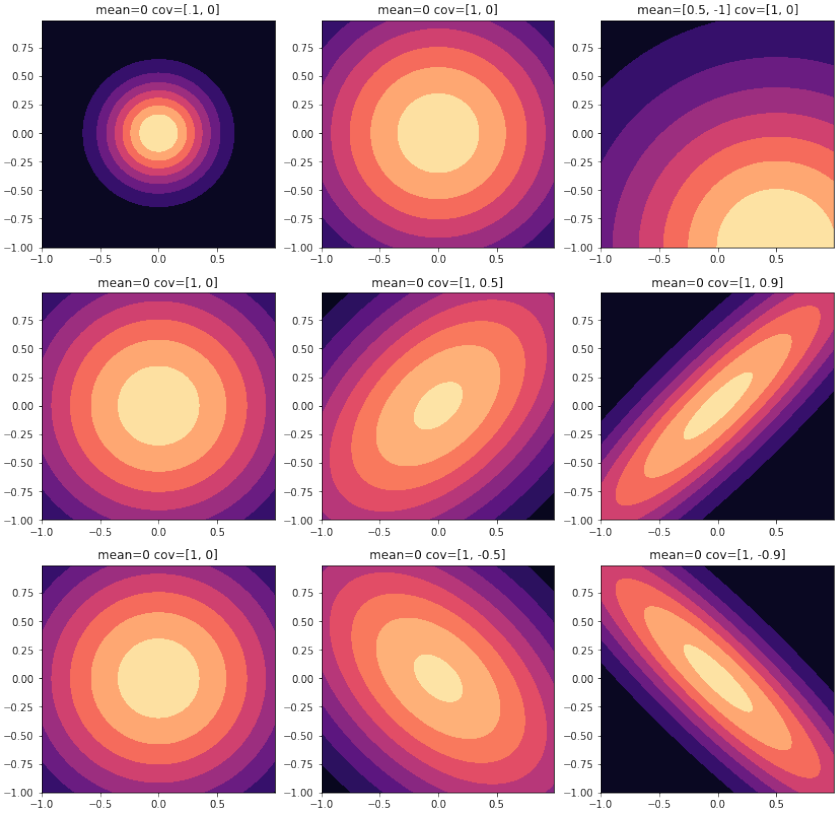

Figura 1: Lugar xeométrico dos puntos con densidade constante para varias distribucións gaussianas con distintas Σ.

- **Clasificador de máxima verosimilitude ($\Sigma_i \neq \Sigma_j \ \forall i \neq j$)** Este clasificador é o máis complexo e xeral do caso normal. As matrices de covarianza son distintas para todas as clases. Isto implica que as funcións discriminantes son cadráticas, polo que as veces se denomina da mesma formas e as superficies son hipercadráticas. Na Fig. 2 amosamos o caso de días clases onde a superficie de decisión $P(\vec{x}|G_i) = P(\vec{x}|G_j)$ é a liña entre os dous grupos que pasa a través da intersección dos contornos equiprobables.

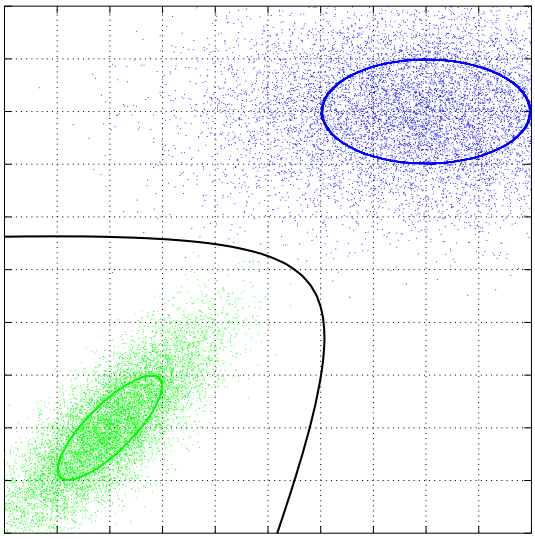

Figura 2: Division cadrática do espazo de características.

A vantaxe deste clasificador é que ten en conta a dispersión da área de cada clase a través da matriz de covarianza o que o converte nun método robusto.

- **Clasificador bayesiano linear ($\Sigma = \Sigma_i = \Sigma_j$)** Aínda que a regra normal de Bayes é unha simplificación do caso xeral, requírese a estimación dos parámetros matriciais (media e covarianza) para cada clase. Sen embargo, aínda se pode simplificar máis se consideramos a mesma matriz de covarianza para todas as clases[2] o que permitiría unha simplificación da función discriminante eliminando termos comúns:

$$d_i(X) = (\vec{x} - \vec{\mu_i})^T \Sigma^{-1} (\vec{x} - \vec{\mu_i}) - \ln(P(G_i)), \tag{16}$$

Esta expresión é lineal, reducíndose o número de parámetros a estimar aumentando a súa eficiencia dende o punto de vista computacional. Nun espazo de características bidimensionais, $d_i(X)$ é unha liña recta (Fig. 3) que divide a dito espazo en dúas rexións. Debido a este carácter linear de $d_i(X)$, este clasificador é coñecido como clasificador bayesiano linear.

Na práctica nunca se cumpre que $\Sigma = \Sigma_i = \Sigma_j$, polo que é preciso un criterio de aproximación. Se embargo, e a pesar do incumprimento das hipóteses que deron lugar a este clasificador, a estimación de $Sigma$ como promedio de todas as matrices de covarianza de cada clase da bos resultados na práctica e as veces superiores ao discriminante cadrático.

[2] Isto darase se as correlacións son independentes en cada clase

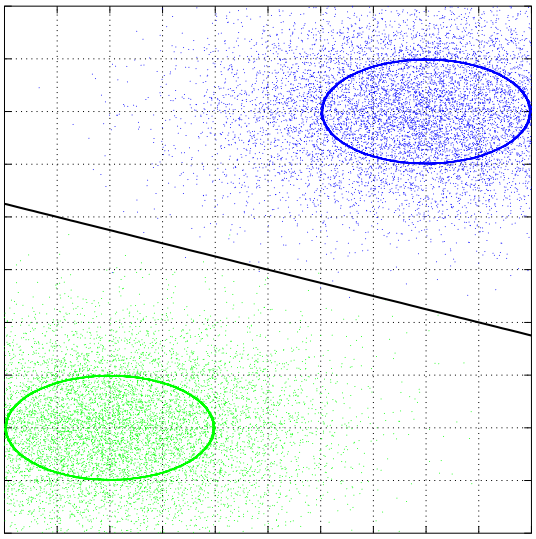

Figura 3: Division linear do espazo de características.

- **Clasificadores de mínima distancia pesada ($\Sigma_i = \sigma^2 I$)**: Este caso ocorre cando as características son estatisticamente independentes e cada unha delas posúe a mesma varianza, σ^2. Neste caso, a matriz de covarianza é diagonal. Xeometricamente, isto corresponde a que as mostras de cada clase están incluídas en conxunto hiperesféricos de igual tamaño e centradas nas súas correspondentes medias, μ_i. Neste caso, a regra normal bayesiana pode ser simplificada, obtendo función discriminante moi simples:

$$d_i(X) = \frac{(\vec{x} - \vec{\mu_i})^T)(\vec{x} - \vec{\mu_i})}{\sigma^2} + \ln(P(G_i)), \tag{17}$$

  No caso de que $P(G_i) = P(G_j)$ e $\sigma = 1$, $\forall i \neq j$, estas probabilidades convértense en constantes que poden ser ignoradas. Neste caso, a clasificación dun vector $\vec{x}$ equivale a determinar a distancia mínima Euclídea deste vector ao centro de cada clase. Este clasificador é coñecido como *mínima distancia*.

4. Medidas de avaliación dos clasificadores

Como dixemos, a clasificación é un tipo de problema de aprendizaxe automática supervisado (no noso caso) onde o obxectivo é predicir, para unha ou varias observacións, a categoría ou clase á que pertencen.

Un elemento importante de calquera fluxo de traballo de aprendizaxe automática é a *avaliación do rendemento* do modelo. Este é o proceso no que usamos o modelo adestrado para facer predicións sobre datos etiquetados non vistos anteriormente. No caso da clasificación, entón avaliamos cantas destas predicións acerta o modelo.

En problemas de clasificación reais, normalmente é imposible que un modelo sexa correcto ao 100 %. Á hora de avaliar un modelo é, polo tanto, útil saber, non só o errado que está o modelo, senón de que xeito erra[3].

Por exemplo, se tratamos de predicir se un tumor é benigno ou maligno, pode ser preferible que o modelo erre en predicir incorrectamente que un tumor é maligno (nun pequeno número de casos) a que prediga que é benigno cando cando non o é con graves consecuencias para o paciente.

Neste caso, optimizaríamos o modelo para que funcione mellor para determinados resultados e, polo tanto, podemos utilizar diferentes métricas para seleccionar o modelo final a empregar. Como consecuencia destas compensacións ao seleccionar un clasificador, hai unha variedade de métricas que debe utilizar para optimizar un modelo para o seu caso de uso específico.

[3] "Todos os modelos son incorrectos, pero algúns son útiles". *George Box*

A continuación, imos a expoñer as oito métricas de rendemento diferentes que se poden usar para avaliar un clasificador.

- **Matriz de confusión** é unha ferramenta extremadamente útil para observar de que xeito o modelo é incorrecto (ou correcto!). É unha matriz que compara o número de predicións para cada clase que son correctas e as incorrectas.

  Nunha matriz de confusión, hai 4 números aos que prestar atención.

  Verdadeiros positivos (TP): o número de observacións positivas que o modelo predixo correctamente como positivas.

  Falsos positivos (FP): o número de observacións negativas que o modelo predixo incorrectamente como positivas.

  Verdadeiros negativos (TN): o número de observacións negativas que o modelo predixo correctamente como negativas.

  Falsos negativos (FN): o número de observacións positivas que o modelo predixo incorrectamente como negativas.

  A Fig. 4 mostra unha matriz de confusión para un clasificador. Usando isto podemos entender o seguinte:

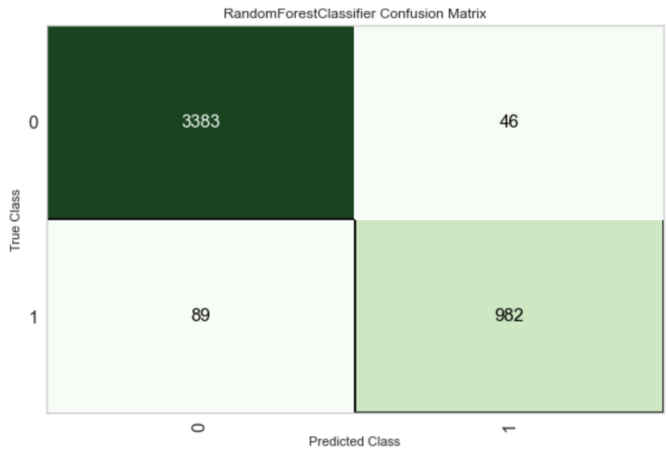

Figura 4: Mariz de confusión.

  O modelo predixo correctamente 3383 mostras negativas pero predixo incorrectamente 46 como positivas. O modelo predixo correctamente 962 observacións positivas pero predixo incorrectamente 89 como negativas. Desta matriz de confusión podemos ver que a mostra de datos está desequilibrada, tendo a clase negativa un maior volume de observacións.

- **Exactitude (Accuracy):** a exactitude global dun modelo é simplemente o número de predicións correctas dividido entre p número total de predicións. Unha puntuación deste medida dará un valor entre 0 e 1, onde o valor de 1 indicaría un modelo perfecto.

$$\mathsf{ACC} = \frac{\#\mathsf{TP} + \#\mathsf{TN}}{\#\mathsf{TP} + \#\mathsf{FP} + \#\mathsf{FN} + \#\mathsf{TN}}. \tag{18}$$

  Se volvemos ao exemplo do cancro. Imaxina que temos un conxunto de datos onde só o 1 % das mostras son cancerígenas. Un clasificador que simplemente prevé todos os resultados como benignos conseguiría unha puntuación de precisión do 99 %. Porén, este modelo sería, de feito, inútil e perigoso xa que nunca detectaría unha observación cancerosa.

- **Precisión (precision)**: mide o bo que é o modelo para identificar correctamente a clase positiva. Noutras palabras, de todas as predicións para a clase positiva, cantas eran realmente correctas?

Usando só esta métrica para optimizar un modelo estaríanos *minimizando os falsos positivos*. Isto podería ser pouco útil para diagnosticar o cancro xa que teriamos pouca comprensión das observacións positivas que non se fan.

$$Pr = \frac{\#\mathsf{TP}}{\#\mathsf{TP} + \#\mathsf{FP}}. \tag{19}$$

- **Sensibilidade (recall) ou True positive rate**: dinos o bo que é o modelo para predicir correctamente *todas as observacións positivas* do conxunto de datos. Non obstante, non inclúe información sobre os falsos positivos polo que sería útil no exemplo do cancro.

  Normalmente, a precisión e sensibilidade obsérvanse xuntos construíndo unha curva de precisión-rememoración. Isto pode axudar a visualizar as compensacións entre as dúas métricas en diferentes limiares.

$$Re = \frac{\#\mathsf{TP}}{\#\mathsf{TP} + \#\mathsf{FN}} \tag{20}$$

- **F1-score**: esta é unha media harmónica da precisión e recall. O valor de F1 será un número entre 0 e 1. Se o valor é 1, indica unha precisión e recall perfectas. Se a puntuación de F1 é 0, isto significa que a precisión ou o recall son 0.

$$F_1 = \frac{2 \cdot \mathsf{recall} \cdot \mathsf{precision}}{\mathsf{recall} + \mathsf{precision}} \tag{21}$$

- **Curva ROC e medida AUC**: un clasificador devolverá a probabilidade de que unha observación pertenza a unha clase particular como resultado da predición. Para que o modelo sexa útil, normalmente convértese nun valor binario, p. ex. ou a mostra pertence á clase ou non. Para iso utilízase un limiar de clasificación, por exemplo, poderiamos dicir que se a probabilidade é superior a $0,5$ entón a mostra pertence á clase 1.

  A curva ROC (Receiver Operating Characteristics) é unha gráfica do rendemento do modelo (unha gráfica da taxa de verdadeiros positivos (TPR) e a taxa de falsos positivos (FPR)) en todos os limiares de clasificación:

$$TPR = Recall = \frac{\#\mathsf{TP}}{\#\mathsf{TP} + \#\mathsf{FN}}, \tag{22}$$

$$FPR = \frac{\#\mathsf{FP}}{\#\mathsf{FP} + \#\mathsf{TN}}. \tag{23}$$

  con estes dúas magnitudes construímos a ROC, como se amosa na Fig. 5

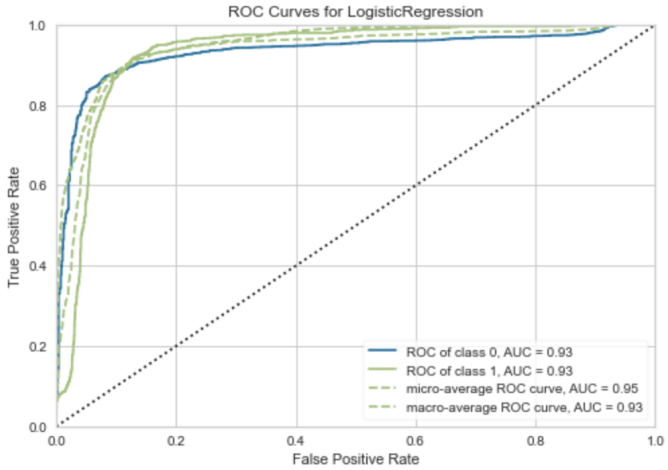

Figura 5: Curva ROC.

A **Area Under Curve (AUC)** é a medida de toda a área bidimensional baixo a curva e, como tal, é unha medida do rendemento do modelo en todos os limiares de clasificación posibles.

As curvas ROC representan a precisión do modelo e, polo tanto, son as máis adecuadas para diagnosticar o rendemento de modelos onde os datos non están desequilibrados.

Que medidas empregar no noso caso? Neste traballo construiremos a matriz de confusión para o número de clases que se usen no noso caso. A partir dela, determinaremos a accuracy xunto coa precision e recall, e a media harmónica destas últimas magnitudes, a $F_1 - score$.

5. Representación e descritores de contorno

5.1. Sinaturas

Unha sinatura define unha función unidimensional obtida a partir dunha forma bidimensional empregando algún método de codificación que reduce a dimensionalidade do problema. A calidade dunha función sinatura mídese en relación as súas características de unicidade, preservación da información e facilidade de cálculo expresada en termos de baixos requirimentos de procesado e memoria. Aínda que se poden pensar múltiples funcións que cumpran cos requirimentos expostos anteriormente, ímonos a centrar nas que empregaremos nesta práctica:

- **Función radial (FR):** esta é a máis intuitiva e directa das dúas que abordaremos. Mediante esta función un contorno caracterízase por unha secuencia ordenada $\{V(k)\}_{k=1,\dots,N}$ na que cada elemento representa a distancia Euclídea entre o centro xeométrico da forma e cada un dos píxeles do contorno.

  Se adoptamos unha representación paramétrica do contorno $\{(x(t), y(t))\}_{t=1,\dots,N}$ entón a función radial a acharemos do seguinte xeito:

$$d(i) = \left[(x(i) - x_c)^2 + (y(i) - y_c)^2\right]^{1/2}. \tag{24}$$

  onde (x_c, y_c) son as coordenadas do centroide da forma.

  O vector $FR = [d_1, d_2, \dots, d_N]^T$ contén todas as distancias dende os puntos do contorno ao centroide e constitúe a nosa función sinatura radial (FR).

  Esta función ten as seguintes propiedades:

- É real e uniavaliada.
- É unha representación global e no dominio espacial.
- Un cambio no punto de comenzo na exploración do contorno. provoca un desprazamento cíclico na sinatura.
- Non é invariante a transformacións de similitude.
- Non preserva a información (pode haber formas con sinatura moi semellante).

- **Sinatura baseada nos ángulos dun triangulo (AT)**

  Sabemos que os ángulos dun triángulo formado por tres puntos (A,B,C) non colineais dun contorno son invariantes baixo calquera transformación de similitude (asumimos que C será o punto sobre o contorno que se estea procesando). Entón, como podemos obter estes tres puntos que precisamos e propiedades deben cumprir? debemos ser capaces de identificar sen ambigüidade dos vértices (A e B) dende o contorno e que sexan minimamente afectados polo ruído. Obviamente, A e B deben ser identificados facilmente. Polo tanto, os mellores candidatos para os vértices A e B son aqueles que poidan ser achados a partir de todos os puntos do borde. Estes puntos poden considéranse como unha extensión dunha forma e non teñen porque estar situados sobre ela. Adicionalmente terían a característica de que son calculados e non detectados (implicaría procesado) e todos os puntos do contorno contribúen colectivamente ao calculo das coordenadas dos puntos específicos. Evidentemente, a función sinatura (AT) derivada con respecto C_{AB} e baseada sobre puntos específicos, ten a propiedade de ser invariante a transformacións de similitude. No noso caso, imos a empregar a elipse fundamental como unha forma específica que nos permitira achar os puntos específicos A e B e o punto de comenzo (a intersección entre o semieixe maior da elipse característica e o contorno). Esta elipse pode ser obterse a partir dos momento xeométricos de segundo orde (ver sección de momentos xeométricos). Se colocamos os puntos característicos (A, B e C) no mesmo plano, imos a ter problemas de descontinuidades. Para solucionar isto, colocaremos vértice A sobre o centroide da forma e o B nun plano perpendicular ao da forma (fóra do plano imaxe). Así, conseguiremos un sistema 3D onde os eixes X e Y se aliñan cos da elipse característica de maneira que a longura entre os vértices A e B é igual ao semieixe maior da elipse característica. Con esta relación, a función sinatura $AT = f(i)$ redefínese como:

$$f(i) = \arctan\left(\frac{C_{AB}}{r(i)}\right) \text{ onde } f(i) \in \left(0, \frac{\pi}{2}\right). \tag{25}$$

onde $C_{AB} = a$ é o semieixe da elipse característica e $r(i)$ é a distancia Euclídea dende un punto do borde ao centro de dita elipse. Así conseguirmo unha función continua para todo punto do contorno.

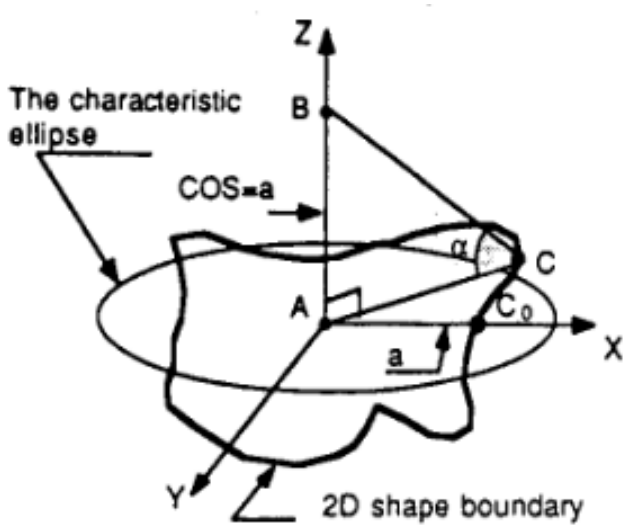

Figura 6: Función sinatura AT.

5.2. Descritores da función sinatura

Os problemas de clasificación poden abordarse directamente a partir da representación ou, indirectamente, mediante algunha descrición de dita representación. Independentemente desta consideración, a forma a empregar a información representativa ou descritiva da lugar aos dous posibles métodos de clasificación, estatístico e estrutural, respectivamente. Optaremos polo primeiro tipo e empregaremos a información en forma dun vector numérico de características. A función sinatura, para todos os puntos do contorno, será unha variable aleatoria e discreta e denotaremos a súa magnitude como $z(i)_{i=1,\dots,N}$, definiremos o noso descritor como:

$$m_r = \frac{1}{N}\sum_{i}^{N}[z(i)]^r. \tag{26}$$

A partir destas magnitudes definiremos o momentos centrais como:

$$M_r = \frac{1}{N}\sum_{i}^{N}[z(i) - \mu_i]^r. \tag{27}$$

Vexamos o comportamento destes descritores fronte a transformación de similitude na función sinatura que pretendemos describir:

- Translación: Un desprazamento na posición dunha forma non ten influencia dado que tanto a sinatura FR como a AT están achadas respecto ao centroide da forma. Polo tanto, os momentos non se verán afectados por esta transformación.
- Rotación: produce un reordenamento circular nos valores das funcións sinaturas (non son invariantes a rotación). Non obstante, o cálculo dos momentos sobre as sinaturas non se ven afectados (non importa o orden dos sumandos).
- Cambio de escala: Para a función FR, se escalamos a forma por un factor α esta función verase afectada como $\alpha z(i)$. Sen embargo, a función angular AT é invariante a cambios de escala porque é un cociente entre a magnitude do eixo maior da elipse característica da forma (varia linealmente coa escala) e a distancia ao centroide (tamén varia linealmente coa escala) producíndose a cancelación do escalado no cociente.

Como consecuencia do anterior, soamente os momentos da función radia FR deben normalizarse a transformacións de escala:

$$M'_r = \frac{1}{N}\sum_{i}^{N}[z'(i) - \mu'_i]^r = \frac{1}{N}\alpha^r\sum_{i}^{N}[z(i) - \mu_i]^r = \alpha^r M_r. \tag{28}$$

Polo tanto, os momentos normalizados poden expresarse como:

$$\hat{M_r} = \frac{M_r}{(M_2)^{r/2}}. \tag{29}$$

e noso vector de características para a función radial como:

$$\left\{\frac{(M_2)^{1/2}}{m_1}, \dots, \frac{M_r}{(M_2)^{r/2}}\right\} \tag{30}$$

con $r > 2$.

5.3. Descritores elípticos de Fourier

Este método permite o cálculo dos descritores de Fourier a partir da descrición dun contorno mediante series elípticas e empregando o código de Freeman de orden oito. As vantaxes deste método resúmense en que non require aproximacións integrais, nin o emprego da FFT, nin unha mostraxe uniforme. Con estes descritores, cun conxunto de dimensión 10 ou menos é suficiente para obter bos resultados para bases de datos cun conxunto de clases suficientemente amplo.

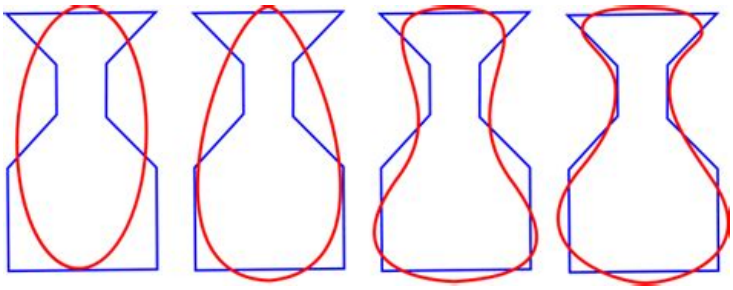

Figura 7: Aproximación á forma orixinal cos descritores elípticos de Fourier da contorna.

Dado que é un pouco longo entrar nos detalles de como se obteñen e normalizan os descritores elípticos de Fourier a partir dus puntos dun contorno, remítese ao estudante ao artigo (aportado) de Kuhl & Giardina[1]. Para obter estes descritores imos a empregar o paquete Pyefd[4]. Nesta URL atoparas como achar os coeficientes, normalizalos e empregalos como descritores en combinación con OpenCV.

6. Descritores de área

A diferencia dos descritores externos, as representacións da rexión veñen dadas por un conxunto de coordenadas cartesianas dos píxeles que pertencen ao obxecto e polo tanto, os algoritmos son máis esixentes computacionalmente que os de contorno. Imos a presentar os que abordaremos neste traballo:

6.1. Momentos xeométricos

Nas tarefas de recoñecemento xeométrico é suficiente empregar imaxes binarias, posto que as propiedades relacionadas coa distribución de intensidades non son moi útiles (si para texturas). Non obstante, as propiedades como área, perímetro, etc. non son axeitadas para problemas reais.

Os descritores de rexión máis utilizados son algún tipo de momentos cos que poden construírse magnitudes invariantes a transformacións de similitude.

Consideremos $f(x, y)$ nun espazo de Hilbert $\mathcal{H}$. Os momentos regulares de orden $(p + q)$ da función $f(x, y)$ están definidos como:

$$m_{pq} = \iint_{\mathbb{R}^2} x^p y^q f(x, y) dx dy, \tag{31}$$

onde $p, q \in N = \{0, 1, ..., \infty\}$. A vista desta definición, pódese considerar que m_{pq} é a proxección da función $f(x, y)$ sobre o monomio $x^p y^q$. Desafortunadamente, a base $\{x^p y^q | p, q \in \mathbb{N}\}$, aínda que e completa en $\mathcal{H}$ non é ortogonal. Se asumimos que $f(x, y)$ é unha función continua a cachos, acoutada e que ten valores distintos de cero só nun parte finita do plano xy, entón existen os momentos de todos os ordes e pode formularse o seguinte teorema de unicidade:

Teorema de unicidade: A secuencia de momentos $\{m_{pq}\}$ está univocamente determinada por $f(x, y)$ e viceversa.

Para o caso de recoñecemento de imaxes esta unicidade cúmprese sempre.

6.1.1. Propiedades dos momentos de orden baixo

Consideremos un plano finito da imaxe na cal se definen as integrais. As propiedades dos momentos de orden baixo son ben coñecidas. Se denotamos por $f(x, y)$ a distribución de radiancia do plano imaxe, o momento de orde cero

[4] pip install pyefd

$$m_{00} = \iint f(x,y)dxdy, \tag{32}$$

representa a masa total da imaxe (ou rexión). Os momentos de primeiro orde

$$m_{10} = \iint xf(x,y)dxdy, \tag{33}$$

$$m_{01} = \iint yf(x,y)dxdy, \tag{34}$$

podemos empregalos para localizar o centroide da rexión da distribución de radiancia,

$$\bar{x} = \frac{m_{10}}{m_{00}}, \quad \bar{y} = \frac{m_{01}}{m_{00}}. \tag{35}$$

Os momentos de segundo orde

$$m_{20} = \iint x^2 f(x,y)dxdy, \tag{36}$$

$$m_{02} = \iint y^2 f(x,y)dxdy, \tag{37}$$

$$m_{11} = \iint xyf(x,y)dxdy, \tag{38}$$

caracterizan o tamaño e a orientación principal da imaxe. De feito, se só se consideran momentos ata o segundo orden, a imaxe orixinal é equivalente a unha función de irradiancia constante e de forma elíptica, tendo definido o seu tamaño, orientación, excentricidade e estando centrada sobre o centroide da distribución de irradiancia. Se asumimos que o centroide é o orixe de coordenadas, os parámetros da imaxe elíptica son,

$$a = \left(\frac{m_{20} + m_{02} + \sqrt{(m_{20} - m_{02})^2 + 4m_{11}^2}}{m_{00}/2} \right)^{1/2}, \tag{39}$$

$$b = \left(\frac{m_{20} + m_{02} - \sqrt{(m_{20} - m_{02})^2 + 4m_{11}^2}^{1/2}}{m_{00}/2} \right)^{1/2}, \tag{40}$$

para os semieixes maior e menor, respectivamente e

$$\theta = \frac{1}{2} \arctan \left(\frac{2m_{11}}{m_{20} - m_{02}} \right), \tag{41}$$

para a orientación da elipse característica.

6.1.2. Momentos centrais e invariantes normalizadas

Os momentos centrais defínense como:

$$\mu_{pq} = \iint_{\mathbb{R}^2} (x - \bar{x})^p (y - \bar{y})^q f(x,y)\, dx\, dy, \tag{42}$$

onde $\bar{y} = \frac{m_{01}}{m_{00}} \bar{y} = \frac{m_{01}}{m_{00}}$ son as coordenadas do centroide.
A fórmula adaptada a imaxes:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x,y), \tag{43}$$

Debido a propia construción, os momentos centrais son invariantes a translación da forma pero non a cambios de escala. Ante unha transformación deste tipo, a nova función de irradiancia é $f'(x,y) =$

$f(x/lambda, y/lambda)$ onde λ é o factor de escala. Baixo esta transformación os cambios nos momentos centrais é,

$$\mu'_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q f(x/lambda, y/lambda) = \lambda^{p+q+2} \mu_{pq}, \tag{44}$$

polo tanto, é evidente que podemos normalizalos ante cambios de escala os momentos centrais da seguinte maneira,

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{(p+q+2)/2}}, \tag{45}$$

ou outra alternativa:

$$\eta_{pq} = \frac{\mu_{pq}}{(\mu_{20}+\mu_{02})^{(p+q+2)/4}}, \tag{46}$$

esta última normalización é máis robusta ao ruído.

Os invariantes de Hu son 7 momentos coñecidos, invariantes á rotación, escala e translación, e calcúlanse a partir dos momentos normais:

$$I_1 = \eta_{20} + \eta_{02} \tag{47}$$

$$I_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{48}$$

$$I_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{49}$$

$$I_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \tag{50}$$

$$I_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \tag{51}$$

$$I_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \tag{52}$$

$$I_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]. \tag{53}$$

$$\tag{54}$$

Despois do traballo de Hu[2], J. Flusser[3] elaborou unha teoría xeral sobre conxuntos completos e independentes de invariantes rotacionais derivados de momentos. Demostrou que o conxunto invariante de Hu non é nin completo nin independente: por unha banda I_3 é redundante, xa que é dependente. noutros, e por outro, falta un invariante, proposto por Flusser e chamado .ºitavo invariante de Hu":

$$I_8 = \eta_{11}[(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2] - (\eta_{20} - \eta_{02})(\eta_{30} + \eta_{12})(\eta_{03} + \eta_{21}). \tag{55}$$

Os momentos de Hu teñen un amplo rango dinámico. Por exemplo, as magnitudes do primeiro momento ($I_1 = 0{,}00162663$) e o sétimo non son comparables ($I_7 = 2{,}09098 \times 10^{-20}$). Por exemplo, podes realizar unha transformación logarítmica da seguinte maneira:

$$I_i = -\text{sign}(h_i) \log |h_i|, \tag{56}$$

Despois de aplicar esta transformación temos para $I_1 = 2{,}78871$ e $I_7 = 19{,}6797$

7. Bases de datos

Nesta práctica, como parte do traballo, os estudantes deben elixir un conxunto (suficiente amplo de clases) de formas 2D (a elixir por cada persoa) como o amosado na Fig. 8.

Figura 8: Exemplo de formas 2D para conformar a base de datos de adestramento dos clasificadores bayesianos.

Este conxunto de formas 2D contén 8 clases de avións, algúns deles bastantes parecidos. Para a adestrar o noso clasificador, precisamos tomar suficientes mostras cunha cámara web de cada clase[5]. Pódese iniciar o adestramento considerando unhas 25 mostras por cada clase e adestrar o noso clasificador con elas e comprobar os resultados.

8. Tarefas

Os estudantes deberán construír dous recoñecedores de formas planas: un con dous tipos de descritores: contorno e área.

Tarefa 1 : Recoñecedores baseados en descritores de contorno

As tarefas que os estudantes teñen que levar a cabo esta parte son:

(a) Busca un conxunto de formas 2D que che interese (sen textura). Emprega unha cámara web para tomar as imaxes e o sistema ten que detectar o contorno e illar a rexión que ocupa a nosa forma no campo visual (estableceremos un limiar e binarizaremos a imaxe). Cando o sistema funcione en produción debe poder recoñecer varias formas que aparezan dentro do seu campo visual (as condicións de iluminación e o fondo quedan a libre decisión do estudante para simplificar a parte de preprocesado e segmentación).

(b) Implementa a función sinatura que che toque (acorde ao número final do teu DNI: se é impar correspóndeche a sinatura **FR** e se é par sinatura **AT**) e obtén os descritores para os contornos que aparecen no campo visual da imaxe. O mesmo para os descritores de Fourier. Cando decidas o número de descritores que empregaras no teu vector de características, visualiza a reconstrucións das formas da base de datos con ese número de descritores.

(c) Cando esteas en modo adestramento, debes ser capaz de gravar os descritores e a súa etiqueta a disco para logo poder adestrar o clasificador.

(d) Unha vez adestrado o teu clasificador, acha as medidas de rendemento: matriz de confusión, accuracy, precision, recall e $F_1 - score$

(e) Presenta os resultados e analízaos.

[5]Con suficientes mostras referímonos a imaxes de cada avión trasladadas a distintos puntos do campo de visión da cámara, rotadas e versión escaladas. Tamén se poden corromper con ruído.

Tarefa 2 : Recoñecedores baseados en momentos xeométricos de área

As tarefas que os estudantes teñen que levar a cabo esta parte son:

(a) A base de datos será a mesma que para os clasificadores de contorno. Debe cumprir cos mesmos requirimentos pero agora para os momentos de área.

(b) Implementa un algoritmo para obter os descritores invariantes de Hu sobre as formas que aparecen no campo de visión. Podes empregar as función de OpenCV que se faciliten o su cálculo.

(c) Cando esteas en modo adestramento, debes ser capaz de gravar os descritores e a súa etiqueta a disco para logo poder adestrar o clasificador.

(d) Unha vez adestrado o teu clasificador, acha as medidas de rendemento: matriz de confusión, accuracy, precision, recall e $F_1 - score$

(e) Presenta os resultados e analízaos. Fai unha comparativa con respecto aos clasificadores de contorno.

Material aportado: código python do clasificador de máxima verosimilitude e as medidas de rendemento para un clasificador. A partir deste programa, podes facilmente programar os demais clasificadores que se piden neste traballo.

9. Entrega

Puntos que debe cumprir a entrega do documento que subas ao Campus Virtual:

1. A práctica debe ser autocontida en dous cadernos de Jupyter: un para recoñecedores baseados en contorno e outra para área.
2. Tabulación dos datos en formato entendible por un humano.
3. Explicación das túas achegas e dos análises dos resultados obtidos.

10. Rúbrica da práctica

- Implementación dos clasificadores bayesianos e medidas simples $\rightarrow$ **25 pts**
- Descritores de contorno e área $\rightarrow$ **25 pts**
- Sistema en produción (python) que recoñeceza e etiquete na imaxe calquera forma da base de datos que entre no campo de visión $\rightarrow$ **25 pts**
- Análise e presentación dos resultados $\rightarrow$ **25 pts**

Referencias

[1] F. P. Kuhl and C. R. Giardina, "Elliptic fourier features of a closed contour," *Comput. Graph. Image Process.*, vol. 18, pp. 236–258, 1982.

[2] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[3] J. Flusser, "On the independence of rotation moment invariants," *Pattern Recognition*, vol. 33, no. 9, pp. 1405–1410, 2000.