Facultad de Matemática, Astronomía y Física y Computación
Universidad Nacional de Córdoba

Ciencia de Datos

Práctico N°7: SVMs, ROC y κ

Problema 1: Estudiar las implementaciones de Support Vector Machines (SVMs) provistas por scikit-learn.
a) Estudiar las diferencias entre los modelos Support Vector Classification (SVC), Linear Support Vector Classification (LinearSVC), Nu-Support Vector Classification y Linear classifiers with SGD (SGDClassifier). ¿Cuales son los kernels disponibles? Destacar los pros y contras de cada modelo.
b) Para la clasificación multiclase identificar cuales modelos implementan el esquema `one-vs-one` y/o el esquema `one-vs-rest`.
c) Estudiar el significado de los parámetros `C`, `nu`, `gamma`, `coef0`, `degree` y `class_weight` y averiguar cuándo se aplican.

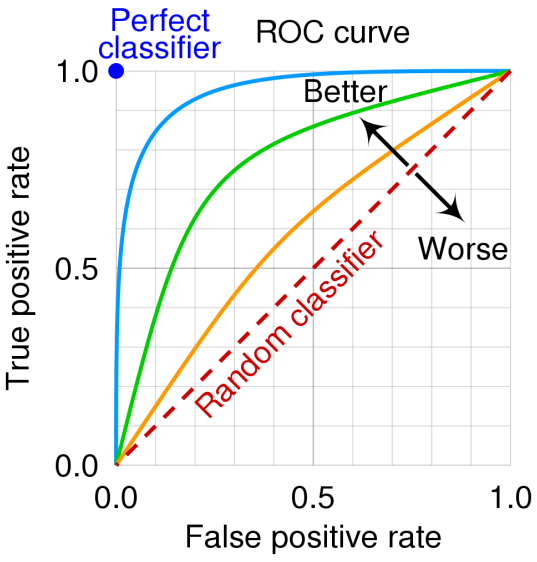
Problema 2: Usar como guía la entrada de scikit-learn sobre SVMs para estudiar las fronteras de decisión generadas por diferentes máquinas de vectores soporte, utilizando como *toy-model* el iris dataset. Para trabajar en el plano 2D, emplear sólo las dimensiones de sépalo de las flores.

Problema 3: Usar como guía la entrada de se scikit-learn sobre RBF SVM para estudiar los parámetros `C` y `gamma`, implementando una búsqueda sobre grid para optimizarlos y visualizar los resultados aplicando el modelo sobre el iris dataset.

Problema 4: En los problemas de clasificación binaria, todas las métricas de performance de un clasificador se construyen a partir de los conceptos básicos TP, TN, FP y FN que constituyen las entradas de la matriz de confusión. Cuando se quiere mostrar la performance de clasificación y se tiene un parámetro disponible en el aprendedor, para cada valor del parámetro se obtiene una matriz de confusión. Una manera de sistematizar toda esa información se consigue representando cada matriz de confusión por el par de valores (FPR, TPR), donde

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}}, \qquad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}},$$

es decir, la razón de falsos negativos y la razón de verdaderos positivos (o recall).
Graficando esos pares de valores en el plano cartesiano, surge la llamada curva Receiver Operating Characteristic (**ROC**). Posibles curvas ROC se muestran en la figura de la derecha, donde la diagonal a trozos se corresponde con la situación de un clasificador completamente al azar (azar insesgado se corresponde con FPR = TPR = 0.5). Para un clasificador perfecto, por el contrario simpre se obtiene FP=0 y TP=P.



Una medida derivada para reportar un único número a partir de la curva ROC, lo constituye el área bajo la curva (AUC).

a) Varios clasificadores proveen como salida un *score* entre 0 y 1 para cada ejemplo, salida que puede interpretarse como una probabilidad y es una medida para para generar un clasificador binario que asigna etiquetas en base a un umbral (usualmente 0.5).

Si la salida del clasificador está por encima del umbral, se etiqueta con p, caso contrario, etiqueta con n. Conceptualmente, puede variarse el umbral desde 1 hasta 0, construir la correspondiente matriz de confusión resultante en cada paso y luego llevar el par de valores (FPR, TPR) al plano para esbozar la curva ROC.

La tabla adjunta muestra 20 ejemplos sintéticos con clase binaria C = {p, n} y el correspondiente *score* asignado por un clasificador hipotético. Construir *a mano* la curva ROC de este ejemplo, usando el procedimiento descripto en el párrafo anterior. Discutir para qué umbral se obtiene la mejor *accuracy*.

	C	Score		C	Score
1	p	0.90	11	p	0.40
2	p	0.80	12	n	0.39
3	n	0.70	13	p	0.38
4	p	0.60	14	n	0.37
5	p	0.55	15	n	0.36
6	p	0.54	16	n	0.35
7	n	0.53	17	p	0.34
8	n	0.52	18	n	0.33
9	p	0.51	19	p	0.30
10	n	0.505	20	n	0.10

Referencia: Tom Fawcett, *An introduction to ROC analysis*, Pattern Recogn. Lett. **27**, 861 (2006).

b) Usando RocCurveDisplay.from_estimator de scikit-learn, graficar la curva ROC de SVC aplicado al Breast cancer dataset.

Problema 5: Un problema usual en el etiquetado de los ejemplos de una base de datos consiste en medir la concordancia entre dos anotadores. En clasificación binaria, la forma más sencilla es calcular la fracción de ejemplos igualmente clasificados. Sin embargo, esta medida no tiene en cuenta las coincidencias por mero azar. Para contemplar esta posibilidad es que se introduce el Coeficiente Kappa de Cohen según,

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

donde p_o es el acuerdo relativo entre los dos clasificadores y p_e es la probabilidad de acuardo hipotético por azar, bajo el supuesto que los clasificadores son independientes. A modo de ejemplo se muestra la matriz de confusión de dos anotadores para 50 ejemplos.

Claramente $p_o = (20 + 15)/50 = 0{,}7$. Por otro lado, uno de los anotadores asigna Yes con probabilidad $(20 + 5)/50 = 0{,}5$ y No con probabilidad $(10 + 15)/50 = 0{,}5$; mientras que el otro asigna Yes con $(20 + 10/50 = 0{,}6$ y No con probabilidad $(5 + 15)/50 = 0{,}4$.

	Yes	No
Yes	20	5
No	10	15

Bajo la hipótesis de independencia: $p_e = 0{,}5 \times 0{,}6 + 0{,}5 \times 0{,}4 = 0{,}5$ y resulta $\kappa = 0{,}4$.

a) Calcular κ en el siguiente ejemplo:

	Yes	No
Yes	25	35
No	5	35

b) Estudiar la implementación de la función κ de scikit-learn y evaluarla en el problema anterior.

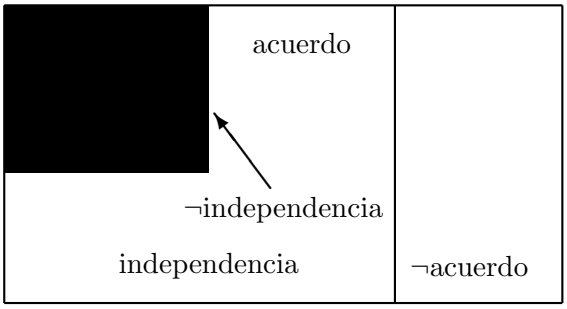
FaMAF 2023

Apéndice: Interpretación probabilística del coeficiente kappa de Cohen

Dados dos clasificadores que rotulan un conjunto de instancias en categorías mutuamente excluyentes, el *coeficiente kappa* es una medida del porcentaje de *acuerdo* entre los clasificadores bajo la condición de *no independencia*.

Desde un punto de vista probabilístico riguroso, el *coeficiente kappa* es la probabilidad condicional de acuerdo entre los clasificadores, dado que las clasificaciones no son independientes entre sí, es decir que están correlacionadas,

$$\kappa = P(\text{acuerdo} \,|\, \text{no independencia}) = \frac{P(\text{acuerdo} \cap \neg\text{independencia})}{P(\neg\text{independencia})} \,. \tag{1}$$



Teniendo en cuenta que el evento (acuerdo), que contiene las instancias en las que los clasificadores coinciden, puede escribirse como unión de los eventos mutuamente excluyentes (acuerdo ∩ independencia) y (acuerdo ∩ ¬independencia) se tiene que

$$P(\text{acuerdo}) = P(\text{acuerdo} \cap \text{independencia}) + P(\text{acuerdo} \cap \neg\text{independencia}) \,. \tag{2}$$

Es importante destacar que la condición de independencia es sólo de interés bajo la condición de acuerdo; i.e., (independencia) ⊂ (acuerdo). Por lo tanto,

$$P(\text{acuerdo} \cap \text{independencia}) = P(\text{independencia}) \,.$$

De esta manera, a partir de la Ec. (1) se obtiene

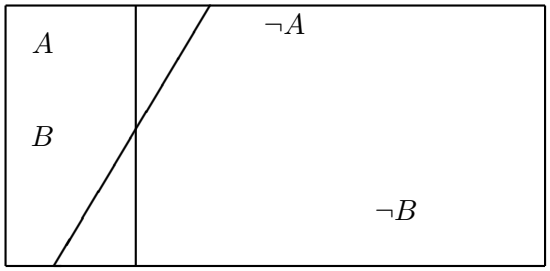
$$\boxed{\kappa = \frac{P(\text{acuerdo}) - P(\text{acuerdo} \cap \text{independencia})}{1 - P(\text{acuerdo} \cap \text{independencia})}} \tag{3}$$

En el caso que los clasificadores sean completamente independientes, es decir,

$$\neg\text{independencia} = \emptyset \quad \text{o bien} \quad \text{acuerdo} = \text{independencia} \,;$$

se tiene que $P(\text{acuerdo} \cap \text{independencia}) = P(\text{acuerdo})$ y resulta $\boxed{\kappa = 0}$. También en el caso trivial (acuerdo) $= \emptyset$ resulta $\kappa = 0$. Por otro lado, si el acuerdo entre los clasificadores es completo, es decir, el evento (acuerdo) se extiende sobre todo el conjunto de instancias; resulta $P(\text{acuerdo}) = 1$ y así $\boxed{\kappa = 1}$. Por último, si los clasificadores están perfectamente correlacionados, (acuerdo) ∩ (independencia) $= \emptyset$; es decir, $P(\text{acuerdo} \cap \text{independencia}) = 0$, resulta $\kappa = P(\text{acuerdo})$.

Para fijar ideas, supongamos dos clasificadores A y B, los cuales rotulan las instancias en dos categorías mutuamente excluyentes (positivo y negativo).



El evento con las instancias rotuladas de igual manera por ambos clasificadores puede escribirse como la unión de dos eventos excluyentes

$$\text{acuerdo} = (A \cap B) \cup (\neg A \cap \neg B) = (A \cap B) \cup \neg(A \cup B)\,. \tag{4}$$

De esta manera,

$$P(\text{acuerdo}) = P(A \cap B) + P(\neg A \cap \neg B) = P(A \cap B) + 1 - P(A \cup B)\,, \tag{5}$$

y usando la condición de independencia,

$$P(\text{acuerdo} \cap \text{independencia}) = P(A)\,P(B) + P(\neg A)\,P(\neg B)\,. \tag{6}$$

Usando,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)\,, \tag{7}$$

y

$$P(\neg A)\,P(\neg B) = (1 - P(A))\,(1 - P(B)) = 1 - P(A) - P(B) + P(A)\,P(B) \tag{8}$$

resulta

$$P(\text{acuerdo}) = 1 - P(A) - P(B) + 2\,P(A \cap B)\,, \tag{9}$$

y

$$P(\text{acuerdo} \cap \text{independencia}) = 1 - P(A) - P(B) + 2\,P(A)\,P(B)\,. \tag{10}$$

De esta forma,

$$P(\text{acuerdo}) - P(\text{acuerdo} \cap \text{independencia}) = 2\,(P(A \cap B) - P(A)\,P(B))\,, \tag{11}$$

y así, finalmente se obtiene

$$\boxed{\kappa = 2\,\frac{P(A \cap B) - P(A)\,P(B)}{P(A) + P(B) - 2\,P(A)\,P(B)}\,.} \tag{12}$$

Puede verse de forma directa que si los clasificadores son independientes, $P(A \cap B) = P(A)\,P(B)$, resulta $\kappa = 0$. Mientras que si el acuerdo entre ellos es perfecto, $A = B$, $P(A \cap B) = P(A)$, entonces $\kappa = 1$.

P.Pury 2012