Tarea 5

Profesor: Felipe Tobar
Auxiliares: N. Aramayo, L. Araya, M. Campos, A. Cuevas y C. Valenzuela
Consultas: C.Valenzuela. (cobavalen AT hotmail.com)
Fecha entrega: 10/5/2018

Formato entrega: Informe en formato PDF, con una extensión máxima de 3 páginas, presentando y analizando sus resultados, y detallando la metodología utilizada. Adicionalmente debe entregar un jupyter notebook (o el código que haya generado) para resolver la tarea.

P1. [Experimento controlado]

El objetivo de este problema es estudiar cómo se comportan diferentes kernels en una situación donde conocemos la clasificación verdadera. Para ello

(**a**) (0.5 puntos) Genere $2n$ puntos (datos X) en $\mathbb{R}^2$ a partir de dos distribuciones normales $\mathcal{N}_0((0,\lambda),I)$ y $\mathcal{N}_1((0,-\lambda),I)$ donde $\lambda > 0$. Debe programar de forma que se pueda probar el código para distintos valores de n y λ.
También debe generar un vector $y \in \{0,1\}^{2n}$ que indique la distribución a la que pertenece cada punto. Grafique dichos puntos asignando distintos colores a cada distribución.

(**b**) (1.5 puntos) Genere $n = 100$ datos y divídalos aleatoriamente en datos de entrenamiento y datos de validación. Escoja un valor de λ tal que los datos sean casi linealmente separables y entrene un clasificador SVM con distintos valores de C (Vea la documentación de `sklearn.svm.SVC`) .
Grafique los errores totales (i.e., datos mal clasificados) que comete su clasificador SVM en función de distintos valores de C, tanto para los datos de entrenamiento como los de validación. Vea para qué valor de C se obtiene el mínimo número de errores tanto en los datos de entrenamiento como en los de validación. Si hay diferencias entre ambos, explíquelas.

(**c**) (1 punto) Usando tres distintos valores de n y tres distintos valores de λ, entrene un SVM con kernel polinomial, rbf y lineal. Defina y calcule una medida de desempeño para estos clasificadores y compare para los valores de n y λ. Considerando que se conoce el clasificador exacto (la asignación de clases ha ido generada sintéticamente), explique sus resultados.

Criterio de evaluación: Esta pregunta busca evaluar su capacidad de generar, preparar y visualizar datos, su entendimiento sobre el algoritmo SVM (en particular el rol del parámetro C) y de cómo interpretar el resultado del algoritmo para diferentes kernels. Todo esto en función de la asignación real de clases.

P2. (3 puntos)

[Datos reales]

Use los datos del archivo *data salaries* disponible en U-Cursos para crear un clasificador que prediga si un individuo gana mas de 50k al año o no. Dichos datos fueron obtenidos de la página de Kaggle,[1] donde también se puede encontrar una descripción de las columnas (la columna fnlwgt **no** es informativa).

Proponga un heurística (y prográmela) para escoger las variables (*features*) que entreguen el mejor desempeño del algoritmo. Justifique detalladamente su estrategia, puede proceder en base a información de la naturaleza

[1] `https://www.kaggle.com/uciml/adult-census-income`

de este problema, de forma completamente libre del problema solo analizando los datos, o bien combinando ambos criterios. Se recomienda comenzar con una elección básica de caraterísticas y luego modificarla

En términos generales, no debiese ser difícil obtener más del 75% de precisión. Es además deseable (pero no obligatorio) que explore y elija valores apropiados para los hiperparámetros del método (en vez de solo confiar en los *por defecto*).

Criterio de evaluación: En general, esta parte buscar evaluar su capacidad para resolver un problema de clasificación con datos reales. En particular, la evaluación tomará en cuenta su elección justificada de :

- Características a utilizar en términos de la naturaleza del problema, como también solo en base a los datos
- Representación numérica para cada característica
- Kernel
- Indicadores de desempeño tanto en los datos de entrenamiento como de validación

Finalmente, la evaluación también considerará su análisis e interpretación de los resultados obtenidos, incluyendo sus ventajas y desventajas.