Curso de aprendizaje automatizado

PCIC, UNAM

Tarea 2: Clasificador bayesiano ingenuo

Fecha límite: 3 de marzo.

Formato: Solución de ejercicio 1 en papel (enviar escaneado) o en archivo electrónico (PDF). Solución de ejercicios 2 y 3 en libretas de Jupyter de manera independiente, reproducible y documentada.

Forma de entrega: Enviar tarea por Google Classroom.

Descripción

Realiza los ejercicios que se describen a continuación y discute ampliamente tus resultados[1].

1. Géneros

Un programa de salud gubernamental desea clasificar los registros de las personas en géneros femenino (F) o masculino (M) a partir de los atributos nombre, estatura y peso. Se cuentan con los siguientes registros:

Nombre	Estatura (m)	Peso (Kg)	Género
Denis	1.72	75.3	M
Guadalupe	1.82	81.6	M
Alex	1.80	86.1	M
Alex	1.70	77.1	M
Cris	1.73	78.2	M
Juan	1.80	74.8	M
Juan	1.80	74.3	M
Denis	1.50	50.5	F
Alex	1.52	45.3	F
Cris	1.62	61.2	F
Rene	1.67	68.0	F
Guadalupe	1.65	58.9	F
Guadalupe	1.75	68.0	F

[1] Todos los ejercicios tienen el mismo peso en la calificación. Es posible usar `scikit-learn` para la solución del segundo y tercer ejercicio pero se espera que al menos uno de los clasificadores se programe únicamente con `NumPy` y `SciPy`.

Entrena un clasificador bayesiano ingenuo usando estimación por máxima verosimilitud y otro usando estimación por máximo a posteriori. Reporta los parámetros que obtuviste en ambos casos y usa los clasificadores entrenados para predecir la clase de los siguientes vectores: $x_1 = (\text{Rene}, 1.68, 65)$, $x_2 = (\text{Guadalupe}, 1.75, 80)$, $x_3 = (\text{Denis}, 1.80, 79)$, $x_4 = (\text{Alex}, 1.90, 85)$ y $x_5 = (\text{Cris}, 1.65, 70)$. Considera un intervalo de $\pm.005$ para el atributo de la estatura y un intervalo de $\pm.05$ para el peso. Describe de forma detallada el procedimiento que seguiste tanto en el entrenamiento como en la predicción y discute los resultados obtenidos.

Para el entrenamiento del clasificador por máximo a posteriori considera los siguientes valores para las distribuciones correspondientes:

Género	Nombre	Estatura			Peso		
	α_k	μ_0	σ_0^2	σ^2	μ_0	σ_0^2	σ^2
M	$2, \forall k$	1.7	0.3	0.0020	85.5	17.0	15.76
F	$2, \forall k$	1.5	0.1	0.0074	70.3	85.0	71.00

2. Spam

Descarga el conjunto de datos de *spam* disponible en `http://turing.iimas.unam.mx/~gibranfp/cursos/aprendizaje_automatizado/data/spam.csv`[2] y realiza lo siguiente:

- Reporta el porcentaje de correos que están etiquetados como *spam* y como *no spam* en el conjunto de datos.
- Divide aleatoriamente el conjunto de datos en el 60 % para entrenamiento, el 20 % para validación y el 20 % restante para prueba usando 0 como semilla para tu generador de números aleatorios.
- Entrena 2 clasificadores bayesianos ingenuos con distintas distribuciones.
- Emplea los clasificadores entrenados para predecir *spam* tanto en los datos de entrenamiento como en los de validación y reporta el porcentaje de predicciones correctas de cada clasificador.
- Discute el desempeño de los diferentes clasificadores
- Reporta el porcentaje de predicciones correctas en el subconjunto de prueba para el clasificador con mejor rendimiento en el subconjunto de validación.

El archivo `spam.csv` contiene 2001 valores por cada renglón, de los cuales los primeros 2000 representan el histograma de palabras de un correo y el último corresponde a la clase, esto es, 1 si es spam y 0 si no lo es.

3. Cáncer de seno

Divide aleatoriamente el conjunto de datos de cáncer de seno de Wisconsin[3] en un subconjunto de entrenamiento con el 60 % de los datos, un subconjunto de validación con el 20 % y un subconjunto

[2] Histogramas de palabras generados a partir de un subconjunto de correos del conjunto de datos de *Enron-Spam*.

[3] el conjunto de datos está disponible en `http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data`

de prueba con el 20 % restante usando 0 como semilla para tu generador de números aleatorios. Este conjunto de datos contiene 699 registros de tumores de seno, de los cuales 458 son benignos y 241 son malignos. Cada registro consta de los siguientes atributos[4]:

Número	**Atributo**	**Valores**
1	Código de la muestra	ID
2	Grosor del tumor	1–10
3	Uniformidad del tamaño de la célula	1–10
4	Uniformidad de la forma de la célula	1–10
5	Adhesión marginal	1–10
6	Tamaño de célula epitelial	1–10
7	Núcleos desnudos	1–10
8	Cromatina blanda	1–10
9	Nucléolos normales	1–10
10	Mitosis de células	1–10
11	Clase	2 para benigno, 4 para maligno

Entrena distintos clasificadores de tumores de seno y evalúalos tanto con el subconjunto de entrenamiento como con el subconjunto de validación y discute su desempeño. Existen 16 registros en el conjunto de datos con un atributo no especificado. Investiga estrategias para rellenar los datos faltantes, utiliza las que consideres más adecuadas para este problema y discute el impacto en el desempeño del clasificador. Reporta el porcentaje de predicciones correctas en el subconjunto de prueba para el clasificador con mejor rendimiento en el subconjunto de validación.

[4]La descripción en inglés se encuentra en `http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names`