Ejemplos Preguntas Aprendizaje Estadístico

Hoja 3 de ejercicios.
Hasta viernes 10 de Noviembre: ejercicios: 9, 10, 11, 13, 14, 15

De la hoja de ejercicios 1:

1. Busca un ejemplo de regresión donde la variable dependiente es cuantitativa. Elabora. ¿Qué es Y? ¿Qué es X? ¿Qué es f(X)? ¿Cómo estimarías f(X) por K nearest neighbours? ¿Cómo se interpretaría esta estimación f(X) en este caso? ¿Cómo lo estimarías con un modelo paramétrico? ¿Cómo se interpretaría esta estimación $\hat{f}(X)$ en este caso?

2. ¿Puedes construir datos de Y y X tal que el error irreducible = 0 y otros tal que es muy alto? ¿Cómo lo representarías en un gráfico? ¿Qué implicaría para la estimación?

3. Busca un ejemplo de un problema de clasificación donde la variable dependiente toma 2 valores. Elabora. ¿Qué es Y? ¿Qué es X? ¿Qué es p(X)? ¿Cómo estimarías p(X) y C(X) en este caso con el método de K nearest neighbours?

4. Busca un ejemplo de un problema de clasificación donde la variable dependiente toma más de 2 valores. Elabora.

5. Un amigo quiere visitarte en Noviembre en Palma. Pregunta por las fechas óptimas y te dice que le gusta el sol. ¿En qué sentido puede verse esto como un problema de clasificación? Elabora. ¿Cómo se podría estimar el problema?

6. Encuentra un ejemplo real de clasificación (simple) donde el error de Bayes es máximo. Y uno donde es mínimo. ¿Qué implica para la estimación?

7. En un problema de clasificación con 3 características describe como son las frontera de decisión de Bayes resultando de una estimación de K nearest neighbours. ¿Para qué sirven? ¿Cómo dependen del K?

Otras Preguntas

8. Explica el problema de "maldición de dimensión". ¿Qué implica?

9. Nos interesa estimar una regresión logística en la que queremos explicar una variable binomial (0,1) con una variable X. Por el problema de endogenidad, lo estimamos en 2 etapas. En la primera etapa en vez de regresar Y directamente sobre X, primero se hace una regresión de X sobre cualquier combinación de instrumentos posibles $\{Z_1, Z_2, Z_3\}$ y se guarda el valor ajustado $\hat{X}$, después se hace una regresión de la variable continua Y sobre $\hat{X}$.

   (a) Explica cómo se utilizaría la validación cruzada para evaluar que combinación de instrumentos da los mejores resultados.

(b) Explica cómo se utilizaría el bootstrap para calcular la desviación estándar del coeficiente de regresión de Y sobre X.

10. Dibuja el resultado de una predicción de 2 vecinos más cercanos de una regresión de los puntos en un examen sobre horas estudiadas. Como te parece la elección de K=2?

11. Analizamos 3 variables con la matriz de correlación dada por

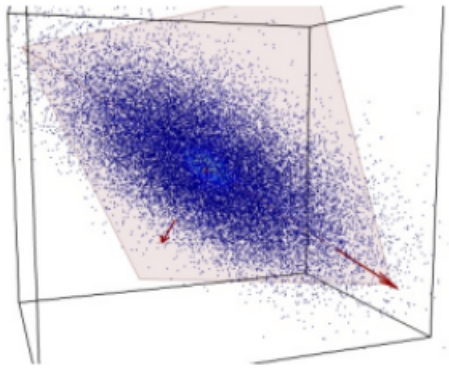
$$\begin{pmatrix} 1 & 0.1 & -0.05 \\ 0.1 & 1 & -0.1 \\ -0.05 & -0.1 & 1 \end{pmatrix}$$

(a) ¿Tiene sentido hacer un análisis de componentes principales?

(b) ¿Cuántos componentes principales crees que encontramos (elegimos)?

(c) ¿Qué porcentaje de la varianza total explica aproximadamente cada componente principal?

12. Explíca el siguiente gráfico en el contexto de un análisis de componentes principales.



13. Te gustaría invertir dinero en acciones. Algún amigo comenta que acciones de tecnologia han crecido los últimos anos más rápidamente. Sin embargo quieres basar tu decisión en datos y comparas los rendimientos de acciones de tecnología con los demás. Encuentras que la diferencia es positiva y altamente significativa. Por eso inviertes. Te parece bien el análisis?

14. Algunos de los resultados de un análisis de componentes principales se muestran en las siguientes tablas. Valore e interprete los resultados obtenidos.

ANALISIS DE COMPONENTES PRINCIPALES SOBRE DATOS DE POLUCIÓN AÉREA

Los datos son observaciones de polución aérea en 80 ciudades americanas (año 1960). El ejemplo proviene de **Jobson, 1992, Applied Multivariate Data Analysis, Springer-Verlag** (data set V7). Las medidas de polución se han obtenido cada dos semanas en las 80 ciudades; las variables se definen como:

SMIN: Nivel mínimo de sulfato (miligramos por metro cúbico x 10).
SMEAN: Media aritmética de las lecturas de los niveles de sulfato (miligramos por metro cúbico x 10).
SMAX: Nivel máximo de sulfato (miligramos por metro cúbico x 10).
PMIN: Cantidad mínima de partículas en el aire (miligramos por metro cúbico x 10).
PMEAN: Media aritmética de partículas suspendidas en el aire (miligramos por metro cúbico x 10).
PMAX: Cantidad máxima de partículas en el aire (miligramos por metro cúbico x 10).

Univariate Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
SMIN	80	47.1000	30.2184	1.0000	155.0000
SMEAN	80	99.6500	50.4276	26.0000	283.0000
SMAX	80	219.8750	120.0390	58.0000	940.0000
PMIN	80	44.5000	18.3806	10.0000	98.0000
PMEAN	80	116.7250	38.8375	54.0000	247.0000
PMAX	80	275.5375	159.0990	117.0000	978.0000

Correlation Matrix

	SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX
SMIN	1.0000	0.5740	0.3024	0.1804	0.1555	-0.0017
SMEAN	0.5740	1.0000	0.8320	0.4481	0.5535	0.3386
SMAX	0.3024	0.8320	1.0000	0.3402	0.5604	0.4738
PMIN	0.1804	0.4481	0.3402	1.0000	0.6951	0.1596
PMEAN	0.1555	0.5535	0.5604	0.6951	1.0000	0.6566
PMAX	-0.0017	0.3386	0.4738	0.1596	0.6566	1.0000

Eigenvalues (CORR)

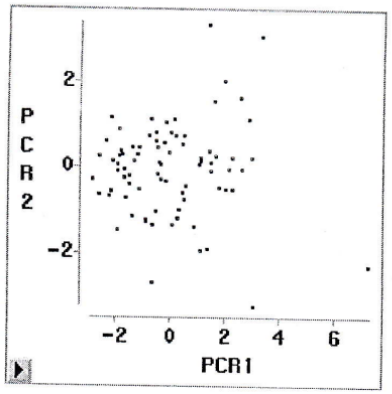
Component	Eigenvalue	Difference	Proportion	Cumulative
PCR1	3.2109	2.0152	0.5352	0.5352
PCR2	1.1957	.	0.1993	0.7344

Eigenvectors (CORR)

Variable	PCR1	PCR2
SMIN	0.2539	0.7090
SMEAN	0.4880	0.3255
SMAX	0.4687	0.0809
PMIN	0.3660	-0.1066
PMEAN	0.4757	-0.3482
PMAX	0.3427	-0.5022

Pattern Matrix (CORR)

Variable	PCR1	PCR2
SMIN	0.4549	0.7753
SMEAN	0.8745	0.3559
SMAX	0.8399	0.0884
PMIN	0.6559	-0.1166
PMEAN	0.8524	-0.3807
PMAX	0.6140	-0.5492



15. Utiliza los datos, en el fichero "satisfacción" y haz un análisis de componentes principales.

FICHERO DATOS: pcegt-satisfacción.sav
Recoge información sobre una encuesta a turistas que han realizado sus vacaciones en Baleares en agosto de 2003. Muestreo por cuotas de nacionalidad realizado en el aeropuerto al finalizar sus vacaciones. Tamaño muestral: 3477.

Se recoge información sobre el nivel de satisfacción global y el nivel de satisfacción de distintas componentes de la oferta turística. También se recoge información de algunas de las características del turista. Las variables de satisfacción se han medido en escala 1 (pésimo) a 10 (excelente).

Variables en el fichero:

SATISGLO	satisfacción global
PAISAJE	satisfacción respecto al paisaje
PLAYA	satisfacción respecto a la playa
CLIMA	satisfacción respecto al clima
CALALOJ	satisfacción respecto a la calidad del alojamiento
CALMEDIO	satisfacción respecto a la calidad medioambiental
CALURBAN	satisfacción respecto a la calidad del entorno urbano
LIMPIEZA	satisfacción respecto a la limpieza
PRCOMID	satisfacción respecto al precio de las comidas
PROCIO	satisfacción respecto al precio de actividades de ocio
PRCOMPRA	satisfacción respecto al precio de compras comerciales
TRATO	satisfacción respecto al trato recibido como cliente
HOSPIT	satisfacción respecto a la hospitalidad de la gente
SEGUR	satisfacción respecto a la seguridad
DIVERS	satisfacción respecto a la diversión nocturna
INFORM	satisfacción respecto a la información
SEÑAL	satisfacción respecto a la señalización
CULTURA	satisfacción respecto a actividades y atractivos culturales
TRANQUIL	satisfacción respecto a la tranquilidad
RUIDO	satisfacción respecto al ruido