CLASIFICACION

**0. La Regla Óptima de Bayes**

La **regla óptima de Bayes** es el clasificador ideal en problemas de clasificación. Asigna una observación X a la clase con la mayor probabilidad posterior P(Y=k∣X). Matemáticamente:



**Ventajas de la Regla de Bayes**

1. **Óptima:** Minimiza la tasa de error de clasificación. Ningún clasificador puede superar su desempeño.
2. **Teórica:** Proporciona un marco ideal para comparar métodos de clasificación.

**Limitaciones**

1. **Desconocimiento de las distribuciones:** En la práctica, no conocemos las distribuciones conjuntas P(X,Y), necesarias para calcular P(Y=k∣X) mediante el teorema de Bayes:

A black and white text

Description automatically generated

1. **Impracticabilidad en datos reales:** Aunque podamos estimar P(X∣Y) y P(Y), la estimación suele ser imprecisa en contextos de alta dimensionalidad o datos limitados. (maldición dimensionalidad)

Por estas razones, los métodos de clasificación práctica, como la regresión logística, LDA o KNN, aproximan de distintas formas la regla de Bayes.

**1. Introducción a la Clasificación**

La clasificación es una tarea de aprendizaje supervisado cuyo objetivo es predecir etiquetas de clase (respuestas cualitativas) basándose en variables predictoras. A diferencia de la regresión, donde el resultado es continuo, la clasificación trabaja con resultados categóricos. Ejemplos comunes incluyen:

* Determinar si un correo electrónico es spam o no spam.
* Clasificar imágenes en categorías como "perro" o "gato".

El proceso de clasificación involucra:

1. **Entrenamiento:** Ajustar un modelo a partir de datos etiquetados (x1,y1),…,(xn,yn)).
2. **Predicción:** Asignar clases a observaciones nuevas (xn+1,xn+2,…).

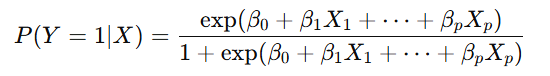
**2. Métodos de Clasificación**

**2.1 Regresión Logística**

La regresión logística modela la probabilidad de que una observación pertenezca a una clase específica, generalmente en problemas de dos clases (Y∈{0,1}).

**Modelo matemático**

El modelo logístico se define como:



Donde:

* P(Y=1∣X): Probabilidad de pertenecer a la clase 1 dado X.
* β0,β1,…,βp​: Coeficientes del modelo que se ajustan a través de máxima verosimilitud.

**Límite de decisión**

Se predice Y=1 si P(Y=1∣X)>0.5, y Y=0 en caso contrario.

**Máxima verosimilitud**

Los coeficientes β se estiman maximizando la función de verosimilitud:

A black and white math equation

Description automatically generated with medium confidence]

**2.2 Modelos Generativos**

Estos modelos asumen una distribución probabilística para las clases Y y para los predictores X dado Y. Los métodos incluyen:

**2.2.1 Análisis Discriminante Lineal (LDA)**

LDA supone que las observaciones en cada clase siguen una distribución normal multivariada con:

* Media específica por clase (μk​).
* Una matriz de covarianza común (Σ\Sigma).

La probabilidad condicional P(Y=k∣X=x) está determinada por el teorema de Bayes:

A number of mathematical symbols

Description automatically generated with medium confidence

Donde:

* fk(x): Densidad normal multivariada de clase k.
* πk​: Probabilidad a priori de la clase k.

La densidad fk(x) se calcula como:

A number of numbers and symbols

Description automatically generated

El clasificador asigna x a la clase que maximiza:

A math equation with numbers and symbols

Description automatically generated

**2.2.2 Análisis Discriminante Cuadrático (QDA)**

Relaja la suposición de matriz de covarianza común. Cada clase k tiene su propia matriz de covarianza (Σk\Sigmak​).

Funciona bien en situaciones donde las varianzas entre clases son diferentes, pero requiere más datos para estimar parámetros adicionales.

**2.3 Naive Bayes**

Asume independencia condicional entre predictores dado Y. La probabilidad posterior P(Y=k∣X) se calcula como:

A black and white text

Description automatically generated

Es eficiente y robusto con muchos predictores, aunque su fuerte suposición de independencia puede ser irrealista en algunos casos.

**2.4 K-Nearest Neighbors (KNN)**

KNN es un método no paramétrico que clasifica observaciones según los k puntos más cercanos en el espacio predictor.

**Pasos del algoritmo**

1. Calcular distancias entre la nueva observación y todas las observaciones de entrenamiento, utilizando una métrica como la distancia Euclidiana: A mathematical equation with a square and a number

   Description automatically generated with medium confidence
2. Identificar los k vecinos más cercanos.
3. Asignar la clase más frecuente entre los vecinos.

**Ventajas y Limitaciones**

* Captura relaciones no lineales entre predictores.
* Sensible a la elección de k y a la escala de los datos.

**3. Evaluación del Desempeño**

**Errores de clasificación**

* **Matriz de confusión:** Muestra el número de predicciones correctas e incorrectas para cada clase.
* **Tasa de error:** Proporción de observaciones mal clasificadas.

**Curva ROC y AUC**

* **ROC (Receiver Operating Characteristic):** Muestra la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR).
* **AUC (Area Under the Curve):** Mide la capacidad del modelo para distinguir entre clases.

**4. Comparación de Métodos**

* **Regresión Logística:** Útil para predicciones probabilísticas. No depende de distribuciones gaussianas.
* **LDA y QDA:** Útiles para clases balanceadas con distribuciones gaussianas.
* **Naive Bayes:** Eficiente con predictores independientes.
* **KNN:** Flexible, pero computacionalmente costoso y sensible a valores atípicos.

**Comparación de Métodos**

Aquí se comparan los principales métodos de clasificación discutidos, agregando **ventajas** y **desventajas** para cada uno.

**Regresión Logística**

* **Ventajas:**
  1. Interpretable: Los coeficientes βj​ tienen una interpretación clara en términos de probabilidades.
  2. Flexible: No asume distribuciones gaussianas para los predictores.
* **Desventajas:**
  1. Linealidad: Supone una relación lineal entre los predictores y el logaritmo de las probabilidades.
  2. Ineficaz para problemas no lineales complejos.

**Análisis Discriminante Lineal (LDA)**

* **Ventajas:**
  1. Eficiencia: Es computacionalmente eficiente y rápido para problemas balanceados.
  2. Robustez: Funciona bien si las clases tienen distribuciones gaussianas.
* **Desventajas:**
  1. Covarianza común: Supone que todas las clases comparten la misma matriz de covarianza, lo cual no siempre es realista.
  2. Problemas en alta dimensionalidad: Se vuelve inestable cuando p>n .

**2.3 Análisis Discriminante Cuadrático (QDA)**

* **Ventajas:**
  1. Flexibilidad: Permite matrices de covarianza diferentes para cada clase, capturando relaciones más complejas.
  2. Útil para datos no lineales: Mejor desempeño en problemas donde las clases no están separadas linealmente.
* **Desventajas:**
  1. Mayor complejidad: Requiere estimar más parámetros (una matriz de covarianza por clase), lo que puede llevar a sobreajuste con pocos datos.
  2. Menor estabilidad: Puede ser menos robusto frente a ruido en los datos.

**2.4 Naive Bayes**

* **Ventajas:**
  1. Eficiencia computacional: Rápido incluso con muchos predictores.
  2. Robustez: Funciona bien con datos de alta dimensionalidad.
* **Desventajas:**
  1. Suposición de independencia: La independencia condicional entre predictores rara vez es válida, lo que puede afectar su desempeño.
  2. Pérdida de información: Ignorar las correlaciones entre predictores puede llevar a clasificaciones subóptimas.

**2.5 K-Nearest Neighbors (KNN)**

* **Ventajas:**
  1. Flexibilidad: No asume ningún modelo subyacente para los datos.
  2. Capacidad no lineal: Funciona bien con fronteras de decisión complejas.
* **Desventajas:**
  1. Costo computacional: El cálculo de distancias para cada nueva observación es intensivo en tiempo y recursos.
  2. Sensible a la escala: Requiere normalizar los datos para evitar sesgos en las distancias.

**Conclusión Comparativa**

1. **LDA:** Rápido y efectivo en clases balanceadas con distribuciones gaussianas.
2. **Regresión Logística:** Útil para interpretar relaciones entre predictores y resultados.
3. **QDA:** Recomendado para clases no lineales y distribuciones heterogéneas.
4. **Naive Bayes:** Simple y eficiente, aunque limitado por la independencia condicional.
5. **KNN:** Excelente para problemas no lineales, pero computacionalmente costoso.

**Fórmulas importantes para recordar**

1. **Límite de decisión en regresión logística:** Predice Y=1 si P(Y=1∣X)>0.5
2. **Verosimilitud en regresión logística:** A math symbols with numbers

   Description automatically generated with medium confidence
3. **Función discriminante de LDA:**

A number and numbers on a white background

Description automatically generated

1. **Distancia en KNN:**

A mathematical equation with a square and a square with a square and a square with a square and a square with a square and a square with a square and a square with a square and a square

Description automatically generated