

Guía de estudio para ETS: Analítica Avanzada de Datos

Enrique Escalante-Notario

Propósito general

Desarrollar sistemas de analítica avanzados de datos con base en Modelos de regresión, clasificación y agrupamiento.

Contenidos:

1. Modelos de regresión avanzados
 - 1.1 Regresión avanzada
 - 1.1.1 Regresión lineal con sustitución no lineal
 - 1.1.2 Regresión robusta
 - 1.1.3 Aplicaciones en la generación de estimaciones de dependencia entre características y aproximadores universales
 - 1.2 Redes avanzadas
 - 1.2.1 Redes neuronales aplicadas a la ciencia de datos
 - 1.2.2 Redes de funciones de base radial
 - 1.2.3 Métodos de validación avanzados
 - 1.2.4 Selección de funciones
 - 1.2.5 Aplicación en la generación de estimaciones de dependencia entre características y aproximadores universales
2. Clasificación
 - 2.1 Criterios de clasificación
 - 2.1.1 Problemas que resuelven las clasificaciones
 - 2.1.2 Características generales de las clasificaciones
 - 2.1.3 Clasificadores a partir del tipo de problema
 - 2.2 Análisis de discriminante lineal y sus algoritmos
 - 2.2.1 Discriminante lineal de Bayes
 - 2.2.2 Discriminante lineal de Fisher
 - 2.3 Máquina de soporte vectorial lineales
 - 2.3.1 Fundamentos
 - 2.3.2 Características
 - 2.3.3 Aplicaciones
 - 2.4 Aprendizaje de la cuantificación vectorial
3. Modelos de agrupamiento en la analítica avanzada
 - 3.1 Agrupamiento basado en prototipos y su empleo en el etiquetado de datos
 - 3.1.1 Hipersférico
 - 3.1.2 Elipsoidal
 - 3.1.3 De formas complejas

- 3.2 Agrupamiento difuso y su empleo en el etiquetado de datos
 - 3.2.1 Partición difusa
 - 3.2.2 Algoritmo Difuso c-Means
 - 3.2.3 Algoritmos posibilístico c-Means
- 3.3 Agrupamiento neuronales difuso y su empleo en el etiquetado de datos
- 3.4 Agrupamiento de aprendizaje competitivo difuso
- 3.5 Agrupamiento difuso adaptivo

1. Modelos de Regresión Avanzados

Recomendaciones de lectura

- **Practical Statistics for Data Scientists** - Bruce & Bruce
 - Capítulo 4: *Regression and Prediction*
 - Fundamentos de regresión lineal.
 - Métodos para manejar datos no lineales mediante transformaciones.
 - Modelos de regresión avanzada, como regresión robusta.
 - Capítulo 6: *Statistical Machine Learning*
 - Modelos predictivos avanzados.
 - Regularización en regresión (Lasso y Ridge).
 - Técnicas para selección de características.
- **Data Analytics: Models and Algorithms for Intelligent Data Analysis** - Runkler
 - Capítulo 6: *Regression*
 - Detalle matemático y práctico de regresión lineal y no lineal.
 - Regresión robusta con diferentes funciones de pérdida.
 - Uso de redes neuronales como aproximadores universales.
 - Ejemplos de aplicación en regresión con datos complejos.

Conceptos clave

- **Regresión lineal con sustitución no lineal** Es una extensión de la regresión lineal tradicional que emplea transformaciones no lineales de las variables independientes para modelar relaciones más complejas entre las variables predictoras y la variable de respuesta.
- **Regresión robusta** Este enfoque minimiza el impacto de valores atípicos en el modelo, utilizando funciones de pérdida alternativas como la de Huber o Tukey para ajustar el modelo.
- **Aproximadores universales** Se refiere a modelos que pueden aproximar cualquier función continua dentro de un dominio cerrado dado un conjunto suficiente de datos. Estos modelos, como las redes neuronales y polinomios no lineales, son útiles para capturar relaciones complejas.
- **Redes neuronales aplicadas a la ciencia de datos** Las redes neuronales son sistemas computacionales inspirados en el cerebro humano, compuestos de capas de nodos (neuronas) que transforman los datos de entrada en información útil para realizar tareas como clasificación y regresión.
- **Redes de funciones de base radial (RBF)** Estas redes utilizan funciones de base radial como activación para resolver problemas de regresión y clasificación, proporcionando una aproximación eficiente y flexible para problemas no lineales.
- **Métodos de validación avanzados** Incluyen técnicas como validación cruzada k-fold y bootstrap, diseñadas para evaluar el rendimiento del modelo de manera confiable y reducir el sobreajuste.
- **Selección de funciones** Es el proceso de identificar y seleccionar las características más relevantes de un conjunto de datos para mejorar la precisión y eficiencia del modelo.

Cuestionario teórico

1. ¿Qué diferencias existen entre un modelo de regresión lineal estándar y uno con sustitución no lineal? Proporcione ejemplos de transformaciones comunes utilizadas para manejar relaciones no lineales.
2. Explique el concepto de regresión robusta. ¿Qué ventaja ofrece frente a la regresión ordinaria cuando se tienen valores atípicos en los datos?
3. Defina el término “aproximador universal”. ¿Por qué se consideran las redes neuronales como aproximadores universales en el contexto de modelos de regresión avanzada?
4. ¿Cuáles son las principales técnicas de regularización utilizadas en modelos de regresión avanzada? Compare Lasso y Ridge en términos de su impacto en los coeficientes de regresión.
5. ¿Qué rol desempeña la selección de funciones en un modelo de regresión? ¿Qué métodos existen para seleccionar las características más relevantes?
6. Describa el proceso de validación cruzada k-fold. ¿Por qué es importante en la evaluación de modelos de regresión avanzada?
7. ¿Qué son las redes de funciones de base radial (RBF) y cómo se diferencian de las redes neuronales tradicionales en aplicaciones de regresión?
8. Proporcione ejemplos prácticos donde la regresión robusta pueda ser más adecuada que otros modelos de regresión avanzada. Explique por qué.
9. ¿Cómo se pueden aplicar los métodos de regresión avanzada para estimar dependencias no lineales entre características en un conjunto de datos?
10. Discuta las ventajas y limitaciones de utilizar modelos de regresión avanzada en problemas de gran escala, como aquellos procesados con herramientas como Spark.

Ejercicios prácticos

1. Regresión lineal con sustitución no lineal:

- Simule un conjunto de datos que relacione la variable independiente x con la variable dependiente y mediante la ecuación $y = 5x^2 + 3x + \epsilon$, donde ϵ es un ruido aleatorio normal con media 0 y desviación estándar 1.
- Ajuste un modelo de regresión lineal estándar a los datos generados y evalúe su rendimiento.
- Aplique una transformación no lineal (e.g., x^2) a los datos y ajuste nuevamente un modelo de regresión. Compare los resultados.

2. Regresión robusta:

- Genere un conjunto de datos lineales $y = 2x + \epsilon$, donde $\epsilon \sim N(0, 1)$.
- Introduzca 5 valores atípicos en los datos simulados.
- Ajuste un modelo de regresión lineal y uno de regresión robusta utilizando una función de pérdida Huber. Compare los coeficientes y evalúe el impacto de los valores atípicos.

3. Aproximadores universales:

- Utilizando una red neuronal con una capa oculta y 10 neuronas, modele la relación entre x y $y = \sin(x)$ en el rango $[0, 2\pi]$.
- Compare el rendimiento de la red neuronal con un modelo polinómico de grado 5.

4. Selección de características:

- Genere un conjunto de datos con 10 variables independientes x_1, x_2, \dots, x_{10} y una variable dependiente y , donde solo x_1, x_2 , y x_3 están relacionadas con y mediante $y = 3x_1 + 2x_2 - x_3 + \epsilon$.

- Aplique un modelo de regresión lineal utilizando todas las variables y luego realice selección de características con regularización Lasso. Compare los modelos obtenidos.

5. Validación cruzada:

- Divida el conjunto de datos del ejercicio anterior en 5 particiones y realice validación cruzada k-fold.
- Reporte el error cuadrático medio (MSE) promedio y evalúe la variabilidad entre las particiones.

6. Redes de funciones de base radial (RBF):

- Genere datos que sigan la función $y = e^{-x^2}$ en el rango $[-3, 3]$.
- Ajuste un modelo de regresión basado en RBF y compare su rendimiento con una red neuronal tradicional con una capa oculta.

7. Aplicaciones prácticas:

- Utilizando el conjunto de datos clásico de California Housing (disponible en bibliotecas como `scikit-learn`), entrene un modelo de regresión robusta para predecir el precio de las viviendas.
- Compare los resultados con un modelo de regresión lineal estándar y evalúe la influencia de los valores atípicos en el rendimiento.

8. Regresión con datos a gran escala:

- Utilizando Spark (o una librería compatible), entrene un modelo de regresión lineal para predecir las llegadas de vuelos con el conjunto de datos de retrasos de vuelos de 2008 (disponible públicamente en Kaggle).
- Realice validación cruzada para evaluar el rendimiento y comente sobre la escalabilidad del modelo.

2. Clasificación

Recomendaciones de lectura

- **Practical Statistics for Data Scientists** - Bruce & Bruce
 - Capítulo 5: *Classification*
 - Introducción a los problemas de clasificación: objetivos y diferencias con modelos de regresión.
 - Tipos de clasificadores:
 - ◊ Clasificadores basados en límites de decisión (e.g., análisis discriminante lineal).
 - ◊ Clasificadores probabilísticos.
 - Métricas de evaluación:
 - ◊ Curva ROC y AUC.
 - ◊ Matriz de confusión y métricas derivadas (precisión, recall, F1).
 - Técnicas avanzadas de clasificación:
 - ◊ Clasificadores combinados, como ensambles.
 - ◊ Discusión sobre sesgo-varianza en clasificación.
 - Capítulo 7: *Unsupervised Learning*
 - Comparación entre agrupamiento y clasificación.
- **Data Analytics: Models and Algorithms for Intelligent Data Analysis** - Runkler
 - Capítulo 8: *Classification Models*

- Modelos de clasificación supervisada:
 - ◊ Descripción del análisis discriminante lineal (LDA) y sus fundamentos matemáticos.
 - ◊ Clasificadores basados en probabilidad, como Naive Bayes.
- Clasificadores basados en árboles de decisión:
 - ◊ Construcción de árboles.
 - ◊ Métricas para decidir divisiones (e.g., ganancia de información).
- Redes neuronales para clasificación:
 - ◊ Introducción al uso de redes neuronales en problemas categóricos.
- Evaluación de clasificadores:
 - ◊ Métricas tradicionales como precisión, recall y F1.
 - ◊ Consideraciones sobre desbalance de clases y su impacto en la evaluación.

Conceptos clave

- **Clasificación supervisada:** Técnica que utiliza datos etiquetados para entrenar un modelo que clasifique nuevas observaciones en categorías predefinidas.
- **Análisis Discriminante Lineal (LDA):** Método lineal que busca maximizar la separación entre las clases proyectando los datos en un espacio de menor dimensión.
- **Clasificadores probabilísticos:** Modelos que predicen la probabilidad de pertenencia de una observación a una clase específica, como Naive Bayes.
- **Árboles de decisión:** Algoritmo basado en una estructura de árbol para tomar decisiones de clasificación, dividido en nodos mediante métricas como ganancia de información.
- **Curva ROC y AUC:** Herramientas gráficas y métricas para evaluar el rendimiento de un clasificador considerando distintos umbrales de decisión.
- **Matriz de confusión:** Representación tabular que muestra las verdaderas clasificaciones frente a las predicciones realizadas por un modelo.
- **Métricas de evaluación:**
 - **Precisión:** Proporción de predicciones correctas sobre el total de predicciones realizadas.
 - **Recall (Sensibilidad):** Capacidad del modelo para identificar correctamente las observaciones positivas.
 - **F1-Score:** Media armónica entre precisión y recall, útil en escenarios de desbalance de clases.
- **Sesgo y varianza:** Conceptos fundamentales para entender el equilibrio entre un modelo que generaliza bien (sesgo bajo) y uno que ajusta perfectamente los datos de entrenamiento (varianza baja).
- **Clasificadores combinados (Ensamblados):** Métodos como Bagging y Boosting que combinan múltiples modelos para mejorar el rendimiento.
- **Desbalance de clases:** Situación en la que las clases no están representadas equitativamente en los datos, lo que puede sesgar los resultados del modelo.

Cuestionario teórico

1. ¿Qué diferencia existe entre un modelo de clasificación supervisada y uno no supervisado? Proporcione ejemplos de cada tipo.
2. Explique el funcionamiento básico del análisis discriminante lineal (LDA). ¿En qué tipo de problemas es más efectivo este método?

3. ¿Cuáles son las principales ventajas y limitaciones de los clasificadores probabilísticos como Naive Bayes?
4. Describa el proceso de construcción de un árbol de decisión. ¿Qué métricas se utilizan comúnmente para decidir las divisiones en los nodos?
5. ¿Qué representan la curva ROC y el área bajo la curva (AUC)? ¿Cómo se interpretan estos indicadores?
6. ¿Qué información proporciona una matriz de confusión? Describa cómo se calculan las métricas de precisión, recall y F1-score a partir de ella.
7. Explique qué es el desbalance de clases y mencione al menos dos estrategias para abordar este problema en los datos.
8. ¿Qué es el sesgo y la varianza en el contexto de los modelos de clasificación? ¿Cómo afectan estos factores el rendimiento del modelo?
9. Compare Bagging y Boosting como técnicas de ensamble. ¿En qué escenarios es más adecuado usar cada una?
10. Proporcione un ejemplo práctico en el que el uso de clasificadores avanzados (como redes neuronales o ensambles) sea más adecuado que métodos más simples, como árboles de decisión.

Ejercicios prácticos

1. Análisis discriminante lineal (LDA):

- Genere un conjunto de datos con dos clases (C_1 y C_2) distribuidas normalmente en un espacio bidimensional con medias $(2, 2)$ y $(-2, -2)$, y una matriz de covarianza $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.
- Aplique LDA para clasificar las observaciones y visualice los límites de decisión.
- Evalúe la precisión del modelo utilizando un conjunto de prueba.

2. Clasificadores probabilísticos (Naive Bayes):

- Utilice el conjunto de datos de ejemplo de Iris (`scikit-learn`).
- Entrene un clasificador Naive Bayes para predecir la especie de una flor basada en las características de longitud y ancho del sépalos y pétalos.
- Evalúe el rendimiento del modelo utilizando una matriz de confusión y reporte métricas como precisión, recall y F1-score.

3. Árboles de decisión:

- Utilizando el conjunto de datos de Titanic (disponible en Kaggle), construya un árbol de decisión para predecir si un pasajero sobrevivió (`Survived`) basado en variables como `Age`, `Fare`, y `Pclass`.
- Visualice el árbol generado y explique las decisiones tomadas en cada nodo.
- Use validación cruzada para evaluar el rendimiento del modelo.

4. Curva ROC y AUC:

- Genere un conjunto de datos con dos clases distribuidas normalmente.
- Entrene un modelo de clasificación binaria (e.g., regresión logística) y calcule la curva ROC.
- Interprete el área bajo la curva (AUC) y su relación con el rendimiento del modelo.

5. Desbalance de clases:

- Simule un conjunto de datos con una clase mayoritaria (90 %) y una clase minoritaria (10 %).
- Entrene un clasificador y evalúe su rendimiento inicial.

- Aplique técnicas para manejar el desbalance, como sobremuestreo de la clase minoritaria o uso de algoritmos como **SMOTE**, y compare los resultados.

6. Técnicas de ensamble:

- Utilizando el conjunto de datos de **Wine** (**scikit-learn**), entrene un modelo de ensamble Bagging (e.g., Random Forest) y uno de Boosting (e.g., Gradient Boosting).
- Compare el rendimiento de ambos modelos utilizando validación cruzada y reporte la precisión promedio.

7. Clasificación con redes neuronales:

- Genere un conjunto de datos con tres clases no linealmente separables utilizando la función **make_moons** de **scikit-learn**.
- Entrene una red neuronal con una capa oculta y active la función ReLU para clasificar las observaciones.
- Evalúe la precisión del modelo y visualice las fronteras de decisión.

3. Modelos de agrupamiento en la analítica avanzada

Recomendaciones de lectura

- **Practical Statistics for Data Scientists** - Bruce & Bruce
 - Capítulo 7: *Unsupervised Learning*
 - Introducción al aprendizaje no supervisado.
 - Fundamentos del agrupamiento como técnica de aprendizaje no supervisado.
 - Algoritmos de agrupamiento clásico:
 - ◊ **k-means**: Descripción y pasos del algoritmo.
 - ◊ **Jerárquico**: Agrupamiento aglomerativo y divisivo.
 - Métricas de evaluación de agrupamiento:
 - ◊ Índices de silueta.
 - ◊ Distancia intragrupo e intergrupo.
- **Data Analytics: Models and Algorithms for Intelligent Data Analysis** - Runkler
 - Capítulo 9: *Clustering*
 - Agrupamiento basado en prototipos:
 - ◊ Hiperesférico (**k-means**) y elipsoidal (**Gaussian Mixture Models**).
 - Métodos de agrupamiento difuso:
 - ◊ Algoritmo **Fuzzy c-means**.
 - ◊ Agrupamiento posibilístico.
 - Agrupamiento jerárquico y su representación mediante dendrogramas.
 - Agrupamiento competitivo y adaptativo.

Conceptos clave

- **Agrupamiento**: Técnica de aprendizaje no supervisado que busca identificar grupos o patrones en los datos basándose en la similitud entre las observaciones.
- **Agrupamiento basado en prototipos**: Métodos que representan cada grupo mediante un prototipo (e.g., el centroide). Ejemplo: **k-means**.
- **k-means**: Algoritmo iterativo que minimiza la varianza intragrupo dividiendo las observaciones en k grupos basados en sus distancias a los centroides.

- **Gaussian Mixture Models (GMM):** Modelo probabilístico que asume que los datos provienen de una mezcla de distribuciones gaussianas, utilizado para agrupamiento elipsoidal.
- **Agrupamiento jerárquico:** Técnica que crea una jerarquía de grupos, representada mediante dendrogramas. Puede ser:
 - **Aglomerativo:** Comienza con cada observación como un grupo y fusiona iterativamente.
 - **Divisivo:** Comienza con un único grupo que se divide iterativamente.
- **Agrupamiento difuso:** Métodos que asignan probabilidades de pertenencia a múltiples grupos en lugar de asignar cada punto a un único grupo. Ejemplo: **Fuzzy c-means**.
- **Fuzzy c-means:** Algoritmo que minimiza una función de costo difusa para encontrar grupos con límites no estrictos.
- **Agrupamiento competitivo:** Métodos en los que los grupos compiten por asignarse observaciones basándose en criterios de similitud.
- **Agrupamiento adaptativo:** Métodos que ajustan dinámicamente el número o las características de los grupos en respuesta a los datos.
- **Métricas de evaluación del agrupamiento:**
 - **Índice de silueta:** Mide la separación y cohesión de los grupos.
 - **Distancia intragrupo e intergrupo:** Evaluación basada en la compactación interna y la separación entre grupos.
- **Aplicaciones de agrupamiento:** Segmentación de clientes, etiquetado de datos no supervisados, identificación de patrones en datos multidimensionales.

Cuestionario teórico

1. ¿Qué es el agrupamiento basado en prototipos? Describa el funcionamiento del algoritmo **k-means** y mencione sus limitaciones.
2. Compare el algoritmo **k-means** con los **Gaussian Mixture Models (GMM)**. ¿En qué escenarios sería preferible utilizar uno sobre el otro?
3. Defina el concepto de agrupamiento jerárquico. Explique la diferencia entre los métodos aglomerativo y divisivo.
4. ¿Qué es el agrupamiento difuso y cómo se diferencia del agrupamiento tradicional? Proporcione un ejemplo práctico donde este enfoque sea útil.
5. Describa el algoritmo **Fuzzy c-means**. ¿Qué ventajas ofrece respecto al agrupamiento tradicional como **k-means**?
6. Explique el concepto de métrica de evaluación de agrupamiento. ¿Qué mide el índice de silueta y cómo se interpreta?
7. ¿Qué son las distancias intragrupo e intergrupo? ¿Cómo afectan estas medidas a la calidad del agrupamiento?
8. ¿En qué consiste el agrupamiento adaptativo? Proporcione un ejemplo donde este enfoque sea necesario.
9. Mencione al menos dos aplicaciones prácticas del agrupamiento en la analítica avanzada. ¿Qué beneficios aporta esta técnica en dichos casos?

Ejercicios prácticos

1. Agrupamiento con *k-means*:

- Genere un conjunto de datos bidimensional con 3 grupos bien definidos utilizando la función `make_blobs` de `scikit-learn`.
- Aplique el algoritmo **k-means** con $k = 3$ y visualice los centroides y las asignaciones de los puntos.
- Experimente con valores diferentes de k y evalúe los resultados utilizando el índice de silueta.

2. Agrupamiento jerárquico:

- Utilizando el conjunto de datos `Iris` (`scikit-learn`), realice un agrupamiento jerárquico aglomerativo.
- Visualice los resultados en un dendrograma.
- Corte el dendrograma en 3 grupos y compare las asignaciones con las etiquetas reales del conjunto de datos.

3. Agrupamiento difuso:

- Genere un conjunto de datos bidimensional con 3 grupos utilizando `make_blobs`.
- Aplique el algoritmo **Fuzzy c-means** y obtenga la probabilidad de pertenencia de cada punto a los diferentes grupos.
- Visualice los grupos formados y comente sobre los puntos con asignaciones difusas (probabilidades cercanas a 0.5 en múltiples grupos).

4. Gaussian Mixture Models (GMM):

- Genere un conjunto de datos bidimensional con grupos elipsoidales utilizando `make_blobs` y añada covarianza entre las dimensiones.
- Aplique el modelo de mezcla gaussiana (**GMM**) para identificar los grupos.
- Compare los resultados del **GMM** con los obtenidos utilizando **k-means**.

5. Evaluación de agrupamiento:

- Utilizando el conjunto de datos generado en el ejercicio anterior, calcule el índice de silueta para los agrupamientos obtenidos con **k-means** y **GMM**.
- Compare los valores y determine qué método proporciona una mejor separación entre grupos.

6. Agrupamiento adaptativo:

- Simule un conjunto de datos con un número desconocido de grupos utilizando `make_blobs`.
- Aplique el algoritmo de agrupamiento adaptativo **DBSCAN** para detectar automáticamente el número de grupos.
- Visualice los resultados y comente sobre la robustez del algoritmo frente a ruido y grupos de formas complejas.

7. Aplicación práctica: Segmentación de clientes:

- Descargue un conjunto de datos de clientes (e.g., `Mall Customer Segmentation Data` disponible en Kaggle).
- Utilice **k-means** para segmentar a los clientes basándose en las características disponibles (e.g., ingreso anual y puntaje de gasto).
- Interprete los grupos obtenidos y proporcione recomendaciones basadas en el análisis.

Referencias

- Bruce, Peter, Bruce, Andrew, and Gedeck, Peter. *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. O'Reilly Media, 2020.
- EMC Education Services. *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley, 2015.
- Nelli, Fabio. *Python Data Analytics*. Apress, 2018.
- Ryza, Sandy, et al. *Advanced Analytics with Spark: Patterns for Learning from Data at Scale*. O'Reilly Media, 2017.
- Runkler, Thomas A. *Data Analytics: Models and Algorithms for Intelligent Data Analysis*. Springer, 2016.