



Estadística Computacional y
Machine Learning
Prof. Jimmy Hernández
Morales

El curso explora las diferentes ramas del Machine Learning y el Aprendizaje Estadístico con un enfoque mixto teórico-práctico. Las técnicas estadísticas que se usan en Machine Learning se estudiarán desde el punto de vista formal. Sin embargo se hará mucho énfasis en la implementación computacional y en las aplicaciones industriales y casos de uso reales. Se usará CPython como lenguaje de programación por ser uno de los lenguajes más utilizados en la industria en el ámbito de ML, pero haremos uso de diferentes bibliotecas, frameworks y software de otros lenguajes como Scala, Java. También se trabajará con motores de bases de datos centrándonos en el uso de NoSQL. NO Seguiremos una bibliografía específica y la mayoría de los temas se darán con el enfoque del profesor con base en su experiencia académica y laboral.

Temario

1. Introducción

- 1.1 Modelo General de Aprendizaje.
- 1.2 Minimización de la función de Riesgo
- 1.3 Patrones de Reconocimiento
- 1.3 Regresión
- 1.3 Descomposición Sesgo-Varianza y Tradeoff
- 1.3 Dimension VC

2. Herramientas Computacionales

- 2.1 Algoritmos
- 2.2 Complejidad en tiempo y en espacio.
- 2.3 Entornos de Desarrollo: Jupyter Notebook
- 2.4 Fuentes de datos y formatos
- 2.5 Introducción al manejo de datos mediante Pandas
- 2.6 Bibliotecas: Numpy, Scipy, Scikit-Learn, Plotly, Matplotlib, pyGAM, entre otros.
- 2.7 Frameworks: Spark, Map-Reduce, TensorFlow y Keras.
- 2.8 Plataformas para Machine Learning y Big Data
- 2.9 Manejador de versiones Git

2.10 Arquitecturas

3. Regresión Lineal y Logística

3.1 Regresión Lineal Simple

3.2 Regresión Múltiple

3.3 Selección de Modelos

3.3.1 Entropía

3.3.3 Entropía relativa o divergencia de Kullback-Leibler

3.3.4 Criterio de Información de Akaike

3.3.4 Criterio de Información Bayesiana

2.3.4 Validación Cruzada (K-Fold Cross Validation)

3.4 Regresión Logística

4. Aprendizaje Supervisado(Clasificación)

4.1 Clasificadores Lineales y Gaussianos.

4.2 Regresión Lineal y Lógica

4.3 Discriminante Lineal(LDA)

4.3 Discriminante Cuadrático (QDA)

4.3 Naive Bayes

4.4 Máquinas de Soporte Vectorial (SVM)

4.5 Kernelización

4.6 Perceptron

4.7 K-NN

4.6 Exactitud, Precisión y Exhaustividad

4.7 Curva ROC y AUC

5. Deep Learning

5.1 Redes Neuronales.

5.2 Forward, Backward propagation

5.3 Introducción a Redes Recurrentes y Redes Convolucionales

5.4 Aplicaciones

5.4.1 Reconocimiento Facial

5.4.1 Segmentación de imágenes

5.4.1 Procesamiento del Lenguaje Natural

6. Árboles y Métodos Ensamble

6.2 Árboles de Regresión

6.3 Árboles de Clasificación

6.5 Bagging

6.6 Métodos Boosting

6.6.1 AddaBost

6.6.2 Gradient Tree Boosting

6.5 Random Forest

7. Modelos Aditivos Generalizados

7.1 Modelos Lineales Generalizados (aspectos generales, el tema es muy amplio)

7.1 Métodos de Suavizado

7.1 Regresión Polinomial Local: Lowess/Loess

7.1 suavizado con Kernels

7.1 Splines

8. Aprendizaje no supervisado(Clustering)

8.1 Reglas de Asociación

8.1.1 Conceptos básicos

8.1.2 Soporte, Confianza y Lift

8.1.2 Algoritmo Apriori

8.2 Análisis de Clúster

8.2.1 Medidas de Similaridad

8.2.2 Métodos basados en particiones

8.2.3 Métodos Jerárquicos

8.3 Métricas de validación internas

8.4.1 Cohesión y separación

8.4.2 Coeficiente de Silhouette

8.4.3 SSW y SSB

8.5 Métricas de validación externas

8.5.1 Entropía

8.5.2 Impureza

8.5.3 Información Mutua

8.6 Análisis de Componentes Principales (PCA)

8.7 Análisis de Componentes independientes (ICA)

9. Machine Learning en ambientes BIG DATA

9.1 Sistemas de Computo Masivos

9.2 API's para la generación de información

9.3 Frameworks Map-Reduce

9.4 Plataformas GeoEspaciales

9.5 Bases de datos NoSql

9.6 Protocolos de Comunicación

9.7 Procesamiento de manera distribuida.

8.8 Paralelización

9.9 Transformaciones y Acciones

9.10 Pipeline, Persistencia y Serialización

10. Estadística de Altas dimensiones

10.1 Descenso por Coordenadas

10.2 Regresión Lasso

10.3 Regresión Ridge

10.4 Endogeneidad Incidental

11. Redes Bayesianas

12. Booststrap

13. Simulated Annealing

Evaluación

Prácticas y Tareas 80 %

2 Exámenes conceptuales: 20 %

Bibliografía

- [1] HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd ed., Springer.
- [2] G. JAMES, D. WITTEN, T. HASTIE AND R. TIBSHIRANI An Introduction to Statistical Learning, with Application in R (Springer, 2013).
- [3] DIGGLE, P Statistical Learning with Sparsity. The lasso and generalizations. Chapman and Hall.
- [4] T. W. ANDERSON An Introduction to Multivariate Statistical Analysis, John Wiley and Sons Inc; Edición: 3rd Edition .
- [5] MICHIE, SPIEGELHALTER Y TAYLOR y PETRICK J. HEAGARTY , Machine Learning, Neural and Statistical Classification.
- [6] ALAN AGRESTI , Categorical Data Analysis, *Wiley-Interscience Publication*.
- [7] THOMAS W. YEE, Vector Generalized Linear and Additive Models with implementations in R, *Springer*.
- [8] ANDREAS MULLER, SARAH GUIDO, Introduction to Machine Learning with Python, a guide for data Scientist *Springer*.