2do parcial Minería de Datos

# Temas:

1. Análisis regresivo multivariado
   1. Regresión polinomial: Cómo determinar el número de monomios y los monomios a incluir? (Algoritmo Genético).
   2. Ver paper “Polynomial Mutivariate Approximation with Genetic Algorithms” (lo envió el profe)
   3. Software “Curve Expert”
   4. Tarea de Wines con Monomios
   5. Preguntarle qué quería de la Tarea
   6. Paper??
      1. Polinomios de Chebyshev (no va venir en el examen)
2. Redes neuronales
   1. Son métodos de regresión multivariada que evitan la maldición de la dimensionalidad. Pero tienen la desventaja de que no se puede interpretar los coeficientes. Se requiere muy poco conocimiento del problema a resolver. “As complex as necessary, as simple as possible”.
   2. Universal Approximation Theory de Cybenko demuestra matemáticamente que si las densidades de entrada son continuas se puede resolver el algoritmo con una sola capa escondida.
   3. Una red neuronal es un filtro adaptativo.
   4. Perceptrón
   5. Paper “The Best Neural Network Architecture”
   6. Backpropagation: Se corrige el error de adelante hacia atrás, el hecho que trabajemos con sigmoides ayuda a que sea derivable y podamos movernos en el gradiente. Este método no garantiza el óptimo global.
   7. Cross-validation
   8. Capas y Neuronas en las capas - si pongo demasiadas neuronas hago overfitting.
      1. Si los datos no son una representación continua se puede usar splines para rellenar o usar un mayor número de capas escondidas.
   9. Cuando detener el entrenamiento? - Early Stopping
   10. Tutoriales Data Engine - Recomiendo el de series de tiempo para predicción de pasajeros para entender:
       1. Learning rate - si es muy chico tarda en converger y puedo quedar atorado en mínimo local, si es muy grande no converge.
       2. Momentum - Backpropagation with momentum - permite ajustar el paso del learning rate en base al gradiente entre el error en este paso y el error anterior.
       3. Decay - hay dos tipos:
          1. Valor de las neuronas va decreciendo, es decir si no le sirve una conexión la red la ignora.
          2. El learning rate, que puede ser dinámico, es decir que el valor de aprendizaje vaya decayendo.
       4. Funciones de activación: (Lineal (capa de input y capa de output), Sigmoidal, Tanh).
       5. Single-Step (meto las observaciones una por una) or Cumulative (meto todo el training set de un jalon)
   11. Cómo usar compresión y entropía para determinar el número de neuronas en la capa oculta
       1. Preproc (PPMZ2): programa para calcular el tamaño efectivo de los datos.
       2. Preproc (NNArchitecture): Nos ayuda a saber el número de neuronas óptimas en la capa intermedia.
3. Algoritmos genéticos
   1. Problema del viajero
   2. No necesita encontrar derivadas, sino que explora el espacio con un enfoque de prueba y error, guardando los mejores resultados.
   3. Cómo funciona un algoritmo genético?:
      1. Definir espacio finito (se puede digitalizar)
      2. Método de codificación (SED: Signo,Enteros,Decimales)
      3. Exploración del espacio de búsqueda de las soluciones
         1. Posibles soluciones.
         2. “Fitness” - evaluas las ecuaciones y “premias” a las que tienen menor error.
            1. Definir una medida de calidad para el fitness
         3. Selección (Cruzamiento?): cuando dos bits funcionan, será replicados más veces
         4. Mutación (modificación pequeña para asegurarnos que la solución óptima está en la población).

Se resuelve el problema de buscar un óptimo desde diferentes puntos simultáneamente.

Si se mantiene fija la función de fitness, desemboca en el óptimo global.

El algoritmo se adapta muy bien a cambios.

1. Máquinas de soporte vectorial (mencionó algo) -- No va venir nada de esto
2. Nearest neighbors
   1. **Paper “ A Novel Cluster Validity Index…” (nos va preguntar algo de aquí) - Ver calidad de los clusters!**
3. Clustering: (nos va preguntar cosas básicas, generales)
   1. Indices de Calidad - Euclidian distance
   2. Cómo determinar el número de clusters? Pregunta sin respuesta determinista pero muchos criterios heurísticos:
      1. Criterio del codo
         1. Información que se va presentando de manera suave y cambia de manera importante (anomalía en el comportamiento informativo.
      2. Otros métodos: Extended penalized competitive learning, GAP statistics, Bayesian Information Criterion, Evolutionary algorithms, Parsimony and homogeneity.
   3. Presentación “Clustering with an N-Dimensional Extension of Gielis Superformula”
   4. Paper “A Clustering Method Based on the Maximum Entropy Principle” (no es necesario leerlo).
   5. *Métodos: K-vecinos más cercanos, Mapas Auto-Organizados (SOM), C-Means (Fuzzy?), Algoritmos Evolutivos, Super-fórmula de Gelis.*
   6. Paper “Clustering of Heterogeneously Typed Data with Soft Computing - A case” Tenemos que leerlo?
   7. Study” Tenemos que leerlo?
   8. Paper “The Search for Irregularly Shaped Clusters in Data Mining” Tenemos que leerlo?
   9. Data Engine Tutorial: Fuzzy clustering for customer segmentation.
      1. Partition Quality (PC - no era calidad pero no recuerdo el nombre)
      2. Partition Entropy (Entropía de las membresías)

# Presentaciones Equipo 2, 3, 4 y 5:

1. Splines
   1. Uso de splines para imputación de datos faltantes:
      1. MAR
      2. MCAR
      3. NMAR
   2. Uso de splines para crecimiento de muestra de datos (generar continuidad en un conjunto de datos para que aplique el teorema de Cybenko que dice que solo se necesita una capa escondida)
   3. Spline cúbico- curvatura mínima
   4. Documento “SPLINES” (lo envió el profe)
2. Teoría de Normalidad
   1. Documento “Sampling for Normality” (lo envió el profe)
   2. Teorema de Límite Central
   3. Teorema de Chebyshev
   4. Pruebas de bondad de ajuste (Chi cuadrada)
3. Muestra Mínima
   1. Cuando usamos una muestra, nos interesa:
      1. Mantener el nivel de información. Usamos Entropía para medir si se mantiene.
      2. Mantener la relación entre las variables. Usamos Algoritmos de Ascenso, Levenberg-Marquardt, Redes Neuronales.
4. Compresión de Datos
   1. Entropía
   2. Teoría estadística de la Información (Shannon)
   3. Teoría algorítmica de la información (Kolmogorov)
   4. Con Pérdida
   5. Sin Pérdida
      1. Huffman (estadístico)
      2. Lempel-Ziv-Welch (diccionario dinamico).
         1. Uso de ventanas deslizantes
   6. Por qué PPMZ2 es superior a Huffman tradicional? Hauffman asumen que los eventos son independientes mientras que PPMZ2 es con probabilidades condicionales.
      1. PPMZ2 - el diccionario no solo contiene el byte inmediato sino el anterior, se requieren 65,556 entradas en el diccionario. El Huffman tradicional solo usa el byte actual.