Parte I
Conceptos

Parte II
Modelos

1. Modelos Lineales

Si disponemos de atributos numéricos y buscamos predecir un atributo también numérico, entonces es natural explorar los modelos lineales. Estos modelos asumen que el atributo a predecir $y^{(k)}$ responde a la forma $w_0 + \sum_i w_i x_i^{(k)}$, es decir a un hiperplano. Como se espera la presencia de errores dentro de las mediciones o simplemente se busca aproximar la forma "real" de y se minimiza alguna métrica de distancia entre $y^{(k)}$ y $y^{(k)'} = w_o + \sum_i w_i x_i$ como pueden ser según [2]:

- Minima Suma de Cuadrados

$$\min \sum_i (y^{(i)} - y^{(i)'})^2$$

- Minimos Valores Absolutos

$$\min \sum_i |y^{(i)} - y^{(i)'}|$$

- M, L y S Estimadores
- Minima Suma Podada de Cuadrados (Least Trimmed Squares)

$$\min \sum_j (y^{(j)} - y^{(j)'})^2, \quad \text{donde} \quad \{y^{(j)} - y^{(j)'}\} \subset \{y^{(i)} - y^{(i)'}\}$$

- Minima Media de Cuadrados

$$\min \text{med}_i (y^{(i)} - y^{(i)'})^2$$

La mayor característica de los métodos lineales es, precisamente, su linealidad. Considerando que es raro encontrar un fenómeno que se comporte linealmente, la linealidad es una desventaja. Sin embargo, esto también permite que el modelo presente un menor sesgo y se verá mas adelante que esto se suele aprovechar.

Segun [3], tenemos que LinearRegression() utiliza la minimización de la suma de los cuadrados sin embargo dispone de un método de selección de atributos basado en el Criterio de Información de Akaike (ver 5) que simplifica el modelo resultante y está activado por defecto.

2. Ejemplos

```
// Cargamos los datos de alguna manera
Instances dataset = ...
// Preparamos una instancia de prueba
Instance testInstance = ...

// Creamos una instancia del modelo
Classifier model = new SimpleLinearRegression();
// Entrenamos con los datos de 'dataset'
model.buildClassifier(dataset);

// Podemos imprimir los detalles del modelo
System.out.println(model);
// Tambien podemos predecir una instancia
double prediction = model.classifyInstance(testInstance);
```

También podemos utilizar LinearRegression(), MultilayerPerceptron()[1] y con algunas modificaciones Logistic().

Parte III
Evaluacion

Parte IV
Selección de Modelos

3. Principio de Parsimonia

Tambien llamado Navaja de Ockham o, en inglés, Ockham's Razor. Es un principio según el cual "entre varias explicaciones equivalentes, se prefiere la mas simple".

Cabe destacar que fueron muchas las críticas contra este principio por "ser imprudente". La *anti navaja* de A. Einstein: "Simple, pero no más simple.".

4. Maximum Log-Likelihood

En español, maximización de verosimilitud. Dado un modelo con parametros θ y un conjunto de muestras X independientes e identicamente distribuidos, la verosimilitud de los parametros dado el conjunto de muestras $\mathcal{L}(\theta|X)$ es igual a la probabilidad de obtener las muestras dado los parametros $P(X|\theta)$. Es decir, cual es la "probabilidad" de que los parametros θ "expliquen" las muestras X.

Luego tenemos que si maximizamos $\mathcal{L}(\theta|X)$ tendremos un conjunto de parametros $\hat{\theta}$ tal que sean o sean lo suficientemente cercanos a los parametros "reales"[2]. Entonces tendremos que $\mathcal{L}(\theta|X) = P(X|\theta) = \prod_i P(x_i|\theta)$, sin embargo maximizar esto en la práctica es dificil por ello se busca maximizar el logaritmo de el, $\text{máx}_\theta\, \mathcal{L}(\theta|X) = \text{máx}_\theta \log \mathcal{L}(\theta|X) = \log \mathcal{L}(\hat{\theta}|X) = \sum_i P(x_i|\theta)$.

[1] Si activamos la opción -*G* (GUI) podremos modificar la topología de la red.

[2] Suponiendo que el modelo "real" sea el mismo que el utilizado.

5. Akaike Information Criterion

El AIC se define como $AIC = 2K - 2\ln\mathcal{L}(\hat{\theta}|X)$, donde K es la cantidad de parametros que se utilizaron en el modelo mas uno y $\mathcal{L}(\hat{\theta}|X)$ es el maximo de la función de verosimilitud. Este criterio indica que el modelo es "mejor" si posee un **menor** valor y es un criterio de comparación entre modelos **bajo los mismos datos**, es decir que el valor en si mismo no es significativo sino si el valor es menor o no al AIC de otro modelo con el cual comparamos. La interpretación es la siguiente: el término $2K$ es la penalización, por tener signo positivo y estamos buscando un menor valor, por la cantidad de parámetros que el modelo posee ya que a mayor cantidad de parámetros **mayor varianza** y mayor posibilidad de "overfitting", por otra parte el término $-2\ln\mathcal{L}(\hat{\theta}|X)$ favorece a una mejor calificación, por el tener un signo negativo, y está relacionado a verosimilitud del modelo para esos parámetros y esas muestras y también la un **menor sesgo**[1].

Referencias

[1] Shuhua Hu. *Estimation error*. 2007.

[2] David Ruppert y Raymond J Carroll. «Trimmed least squares estimation in the linear model». En: *Journal of the American Statistical Association* 75.372 (1980), págs. 828-838.

[3] I.H. Witten, E. Frank y M.A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011. ISBN: 9780080890364. URL: `https://books.google.com.py/books?id=bDtLM8CODsQC`.