Trabajo Final Minería de Texto: ELiC

José M. Castaño

Ejercicios sobre clasificadores

- Hacer solo las partes a) y b) del archivo adjunto (probclas.pdf).
- Las c) y d) son opcionales para credito extra.
- Hacer el ejercicio 13.6, partes i) y ii) (archivo 13.6.pdf)
- Vecinos Más Cercanos

(a) kNN

Considere el siguiente corpus supervisado de titulares de noticias, donde la primera palabra es la clase del documento.

WorldNews	Iraq election
WorldNews	French executive injured
Business	Chief executive smiles
Business	Krispy Kreme executive resigns

(i) Asignele una clase al siguiente documento (el titular de 2 palabras) usando un clasificador 3NN

executive suite

A que clase es asignado? Asuma frecuencia de palabras en bruto ('raw'), sin idf, ni similaridad del coseno. Justifique su respuesta.

(ii) Se obtendría necesariamente el mismo resultado , usando un clasificador 1NN? Justifique el por qué de su respuesta.

(b) Naïve Bayes

(i) Se ha observado que los clasificadores "Naive Bayes" al asumir independencia pueden contar dos veces la misma evidencia. Muestre con un ejemplo como esta situación puede resultar en una decision de clasificación errónea. Base su ejemplo en una collección de textos que contiene el nombre "Mariah Carey" varias veces, pero donde los terminos "Mariah" y "Carey" nunca ocurren si no es juntos como mencionamos.

(ii) Por otra parte, es frecuente asignar mayor peso a zonas en un documento para mejorar la performance de la clasificación. Es esto un forma de "double count"? Explique la diferencia entre una y otra.

- Enviar su trabajo final a: jcastano@dc.uba.ar