uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE
APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS
GRADO EN ESTADÍSTICA Y EMPRESA
Ricardo Aler

Clasificación con muestras desbalanceadas.

Curvas ROC

Contenidos :

- Primero se discute la importancia de tratar con conjuntos de datos desbalanceados (aquellos en los que una de las clases, minoritarias, tiene muchos menos datos que otras, mayoritarias). Esto hace que los métodos de aprendizaje se centren en la clase mayoritaria y tengan tasas de acierto pequeñas para la minoritaria. Estos problemas son muy habituales, en dominios como por ejemplo, medicina, o detección de anomalías.
- Se distingue entre evaluación y aprendizaje
- Se comienza mostrando como se puede utilizar la matriz de confusión para evaluar de manera más precisa los resultados en problemas de muestra desbalanceada, dado que usar la tasa de aciertos (Accuracy) es engañosa, puesto que describe principalmente los resultados de la clase mayoritaria, ignorando la minoritaria.
- Conceptos asociados a la matriz de confusión son True Positives, True Negatives, False Positives, False Negatives. Y los valores normalizados correspondientes (TPR, TNR, FPR, FNR).
- Otras medidas que tienen en cuenta el desbalanceo a la hora de evaluar son la balanced accuracy y la medida F1.
- Aprender o entrenar modelos con muestras desbalanceadas significa que los modelos funcionen bien para todas las clases, no sólo para la mayoritaria.
- Las maneras habituales de forzar a que el método de aprendizaje se centre en la clase minoritaria son : undersampling/oversampling y thresholding.
- Una técnica de sobremuestreo (oversampling) que funciona particularmente bien en estos problemas es SMOTE, la cual no se limita a repetir instancias de la clase minoritaria, sino que crea nuevas muestras sintéticas de dicha clase en posiciones razonables, intercalando las nuevas muestras entre parejas de muestras originales (de la clase minoritaria).

- Para entender la técnica de thresholding, es necesario conocer primero la curva ROC, la cual muestra como cambian los TPR en función de los FPR, a medida que vamos cambiando el threshold (o valor de corte) de un scoring classifier.
- Un scoring classifier es un clasificador que no se limita a informar de la clase predicha, sino que proporciona un score que muestra la seguridad del clasificador en la predicción. El score puede ser una probabilidad (aunque no necesariamente).
- Thresholding significa el valor de threshold adecuado para maximizar alguna medida (por ejemplo, el TPR o la balanced accuracy).
- Por último, la curva ROC permite introducir otra medida adecuada para medir el buen funcionamiento de un clasificador con muestras desbalanceadas : el área bajo la curva ROC (o AUC o AUROC), además de las ya conocidas balanced accuracy o F1.

Material asociado

Además de las diapositivas de la clase, hay un tutorial, y una práctica sobre este tema.