dcc
CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

Curso DM
Clasificación
(Evaluación II)

Primavera 2023

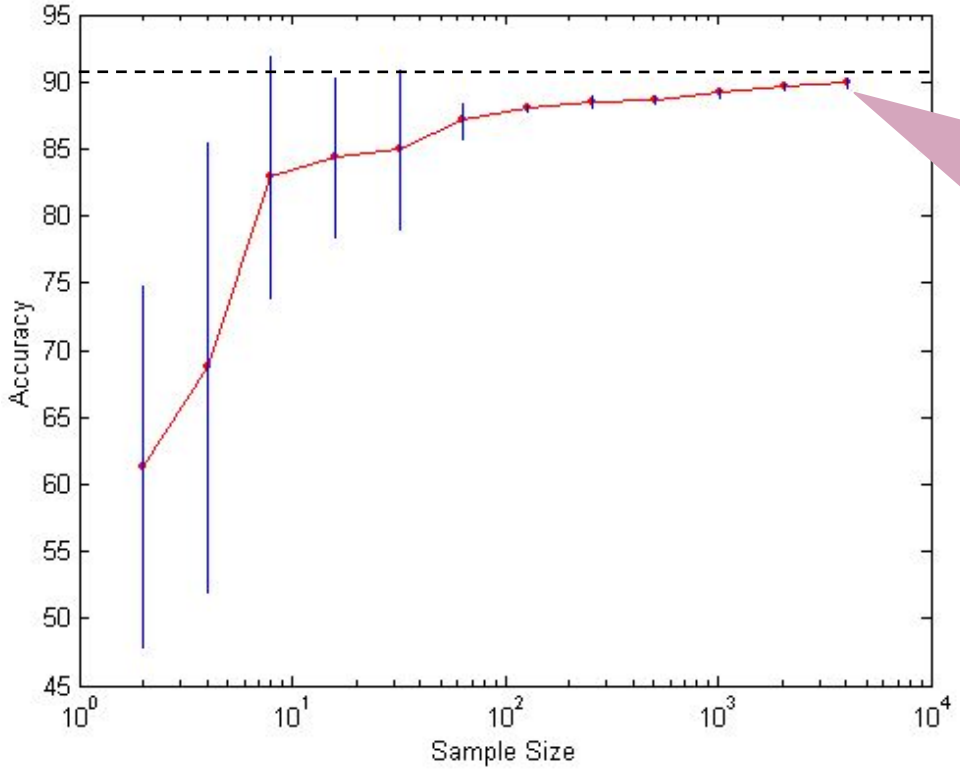
Basado en las slides de Bárbara Poblete

Evaluación del desempeño del modelo

El desempeño de un modelo puede depender de factores diferentes al algoritmo de aprendizaje:

- Distribución de las clases
- Costo de clasificaciones erróneas
- Tamaño de los datos de entrenamiento y test

Curva de aprendizaje



El algún punto, las métricas de nuestro modelo convergen, dado que aumentar la cantidad de datos no genera cambios

Métodos para evaluar el desempeño de un modelo

La idea es estimar la capacidad de generalización de modelo, evaluándolo en datos distintos a los de entrenamiento.

- Holdout
- Random subsampling (subsampleo aleatorio)
- Cross validation (validación cruzada)

Holdout

Particionamos los datos etiquetados en una partición de training y otra de testing.

- Usualmente usamos 2/3 para entrenamiento y 1/3 para evaluación.

Limitaciones:

- La evaluación puede variar mucho según las particiones escogidas.
- Training muy pequeño => modelo sesgado.
- Testing muy pequeño => accuracy poco confiable.

Random Subsampling

Se repite el método holdout varias veces sobre varias particiones de training y testing.

Permite obtener una distribución de los errores o medidas de desempeño.

Limitaciones:

- Puede que algunos datos nunca se usen para entrenar.
- Puede que algunos datos nunca se usen para evaluar.

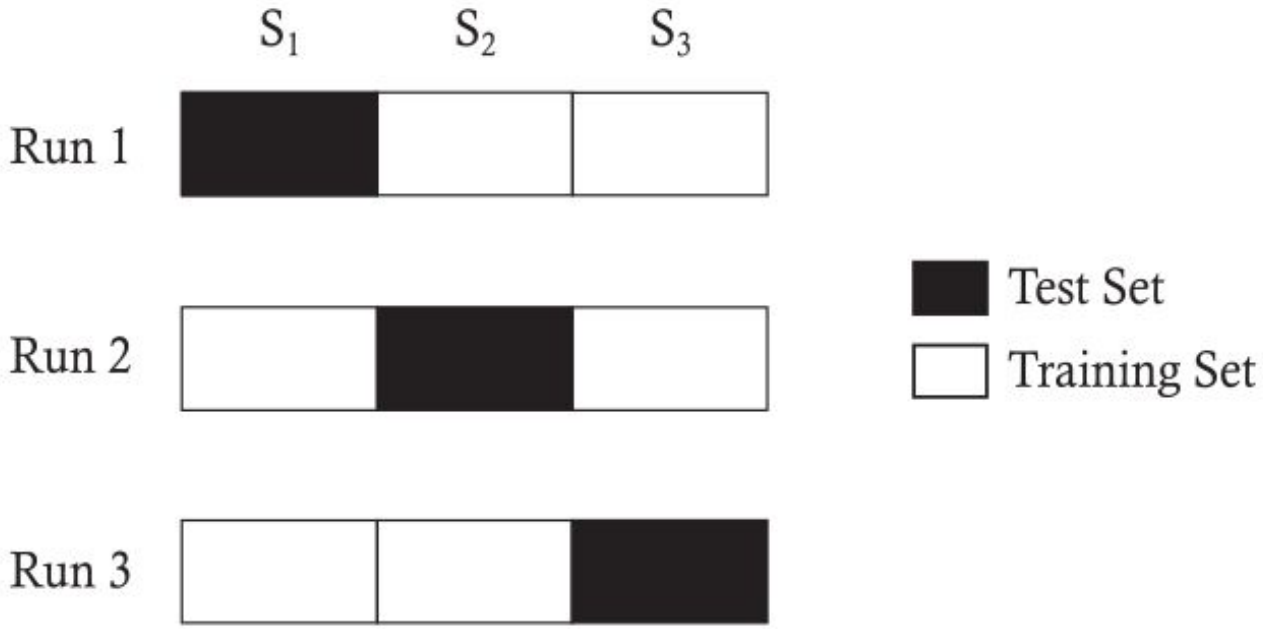
Validación cruzada (cross-validation)

Se particiona el dataset en en k conjuntos disjuntos o folds (manteniendo distribución de las clases en cada fold).

Para cada partición i:

- Juntar todas las k-1 particiones restantes y entrenar el modelo sobre esos datos.
- Evaluar el modelo en la partición i.
- El error total se calcula sumando los errores hechos en cada fold de testing.
- Estamos entrenando el modelo k veces.
- Variante: leave-one-out (k=n)

Validación cruzada (cross-validation)

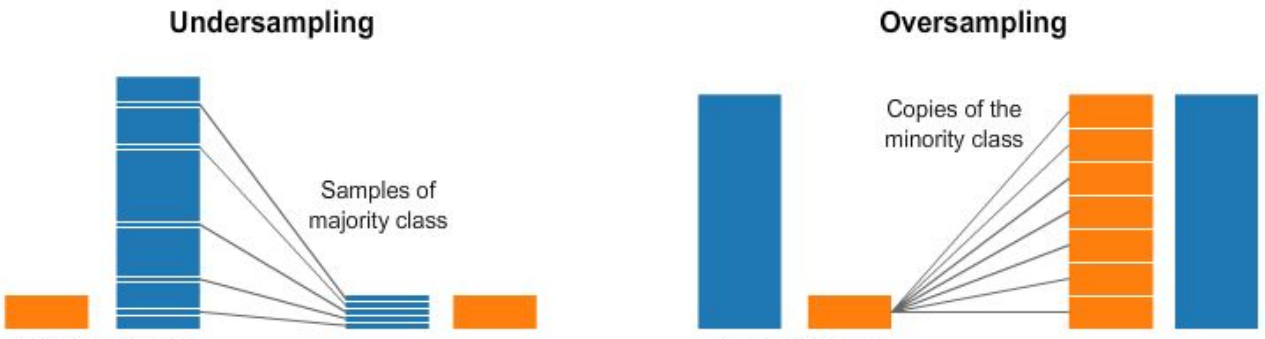


Trabajar con clases desbalanceadas

Para mejorar el rendimiento de un clasificador cuando se tienen clases desbalanceadas existen varias técnicas. Por ejemplo:

Oversampling: Repetir aleatoriamente ejemplos de la clase minoritaria.

Undersampling: Eliminar aleatoriamente ejemplos de la clase mayoritaria.



Fuente: https://www.kaggle.com/code/rafjaa/resampling-strategies-for-imbalanced-datasets/notebook

Trabajar con clases desbalanceadas

- Antes de hacer algo para tratar el desbalance entre las clases primero debemos dividir en train-test.
- Aplicar oversampling y/o subsampling **únicamente sobre la partición de entrenamiento (train)**.
    - ¡Precaución! Si se aplicase (erroneamente) a todo el dataset, el test no será una fiel representación de lo que ocurre en realidad.

Problemas prácticos en la clasificación

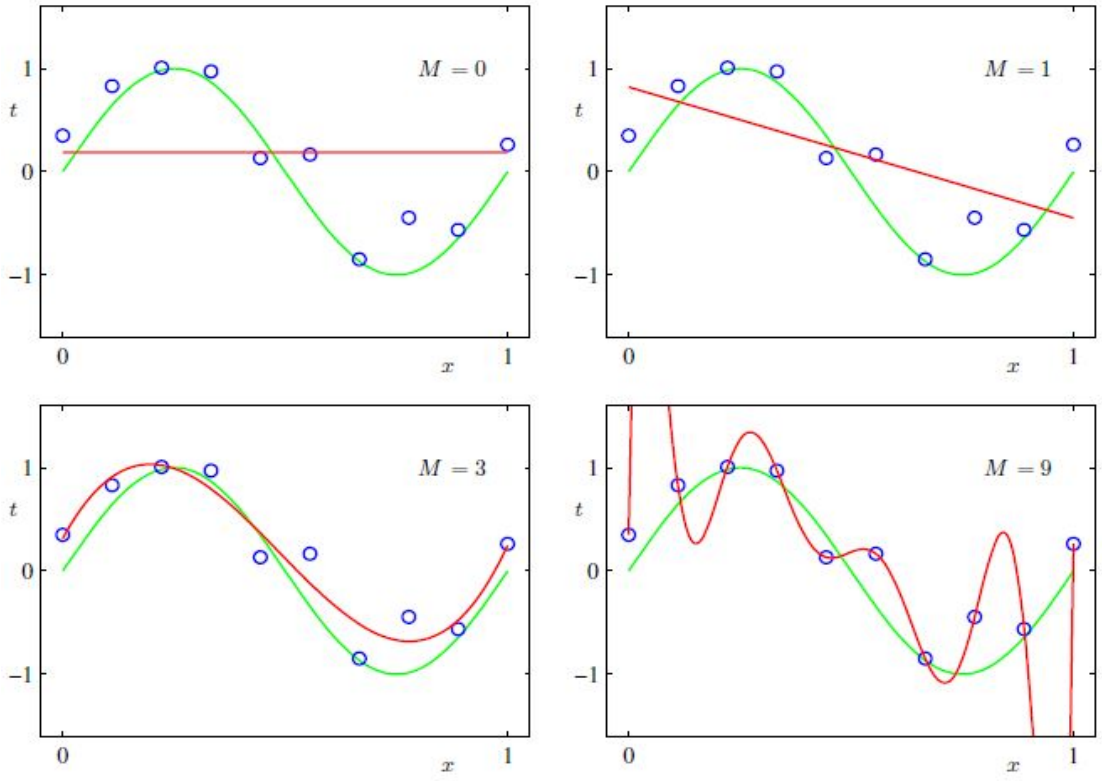
- **Errores de entrenamiento** (malos resultados sobre los datos de entrenamiento): esto ocurre cuando el clasificador no tiene capacidad de aprender el patrón.

- **Errores de generalización** (malos resultados sobre datos nuevos): esto ocurre cuando el modelo se hace demasiado específico a los datos de entrenamiento.

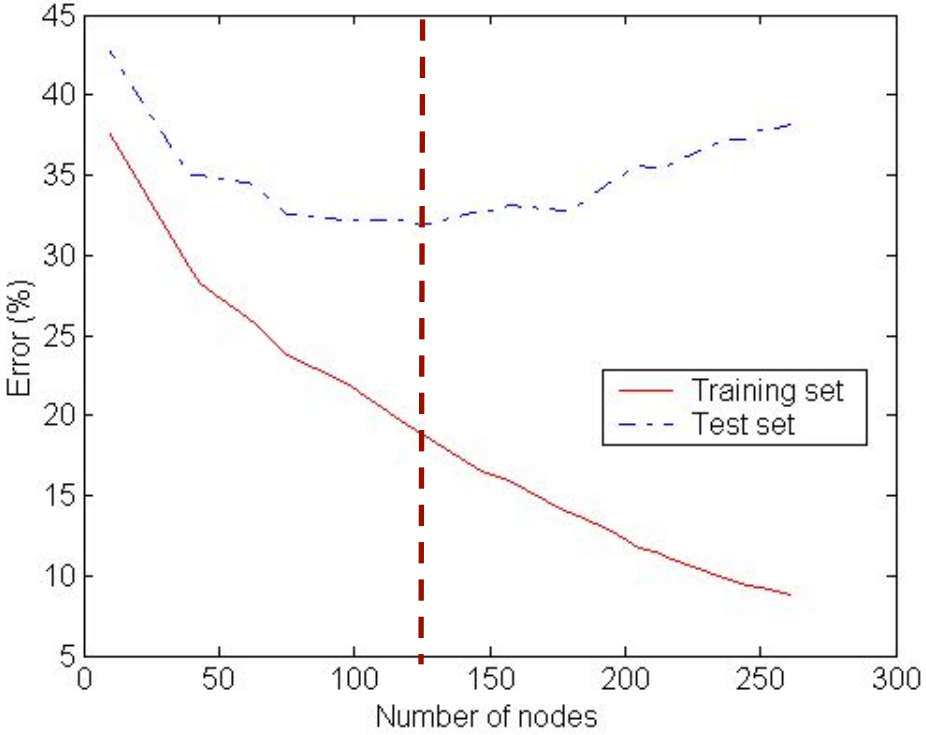
Ambos tipos errores deben ser bajos en un buen modelo

Overfitting y Underfitting

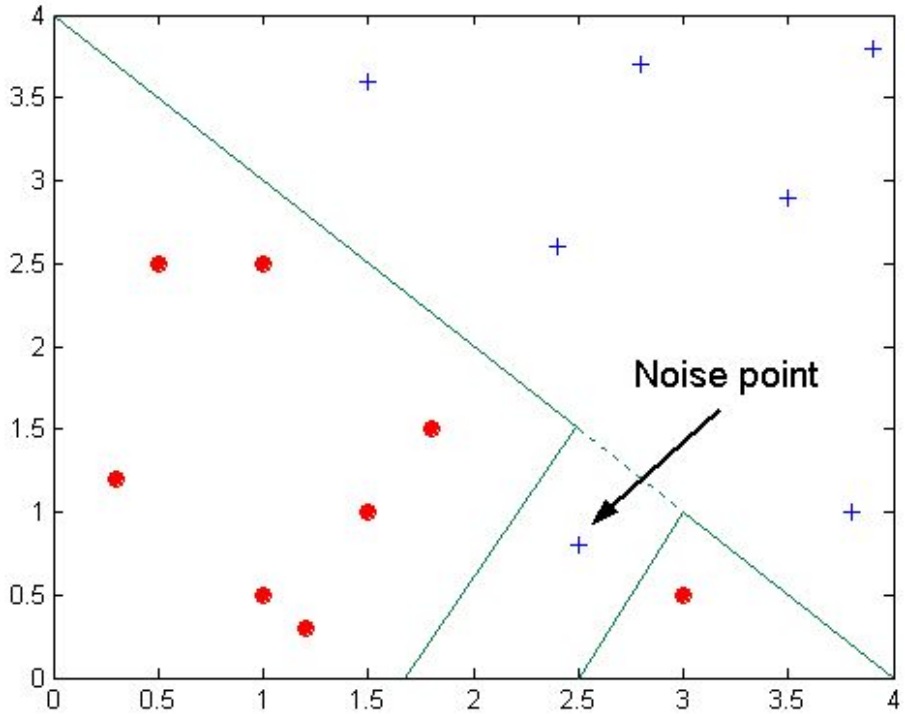
usando polinomios para un problema de regresión



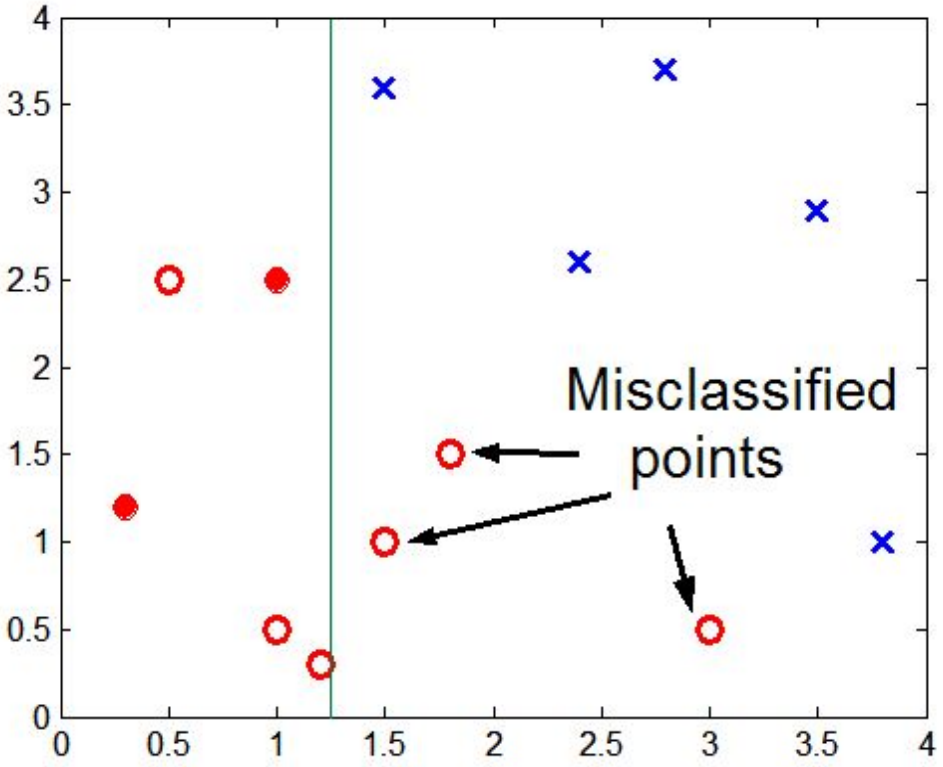
Overfitting y Underfitting



Overfitting por ruido



Overfitting por ejemplos insuficientes



Notas sobre el Overfitting

- El overfitting es un reflejo de un modelo más complejo que lo necesario.

- El error de entrenamiento no es un indicador confiable de cómo se desempeñaría el modelo sobre datos nuevos.

Curva ROC
(Receiver Operating Characteristic Curve)

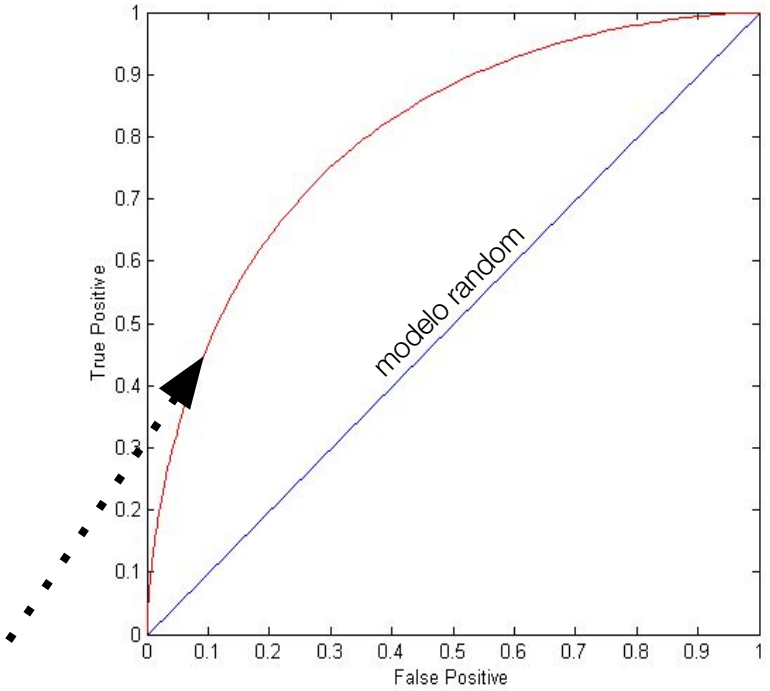
- De manera similar que el trade-off entre Precision y Recall también existe un tradeoff entre la tasa de verdaderos positivos y la tasa de falsos positivos.

  TP Rate: TP / (TP + FN)
  FP Rate: FP / (FP + TN)

- La curva ROC se construye graficando TP Rate vs FP Rate para varios umbrales de clasificación de un clasificador probabilístico (ej: regresión logística, naive Bayes).

Curva ROC
(Receiver Operating Characteristic Curve)

- Entre mayor sea el área bajo la curva mejor es el modelo.
- El área bajo la curva ROC se conoce como AUC y es una métrica ampliamente usada.
- Un tutorial recomendado: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5



dcc

CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

www.dcc.uchile.cl