1. Introducción

Que es el aprendizaje automático?

Relacionar automática mente la entrada de un sistema con su respuesta.

Que es el hold-out y la validación cruzada?

Métodos utilizados para cuantificar la calidad de un clasificador supervisado.

Holt-out: Seperar los datos en training + test
Validación cruzada (k-fold): generar k conjuntos de datos basado en fragmentos del dataset completo, siempre dejando un fragmente fuera para testing. Se entrenan k modelos y se hace el testing con el fragmento de datos que se dejo fuera.
La media de este entrenamiento sirve como aproximación para saber que tal el algoritmo esta generalizando.

2. Evaluación de Algoritmos de regresión

Que es una regresión?

La regresión es un modelo supervisado que relaciona unos variables predictores con una variable respuesta.

Hay dos tipos de errores:
Error reducible: error por falte de datos de entrada
Error irreducible: error por ruido/características estocásticas del proceso

Métricas de error

- MSE (mean square error)
- MAE (mean absolute error): cada error contribuye igual
- RMSE (root mean square error): amplifica errores grandes
- RMLSE (root mean logarithmic square error): penalize underprediction

Visualizar los resultados + correlación

- Visualizar la regresión con un plot scatter (x = predicción, y = respuesta)
- Calcular la correlación entre predicción y respuesta

Que es la 'curse of dimensionality'?

La distancia entre puntos crece con cada dimensión añadida, así que en el espacio de alta dimensión cada datset esta sparse.

3. Evaluación de algoritmos de clasificación

Que es la matriz de confusión y que son valores calculada de ella?

La matriz de confusión se utiliza para evaluar la calidad de un **clasificador**.
Tiene los ejes: Valor de Verdad vs. Predición

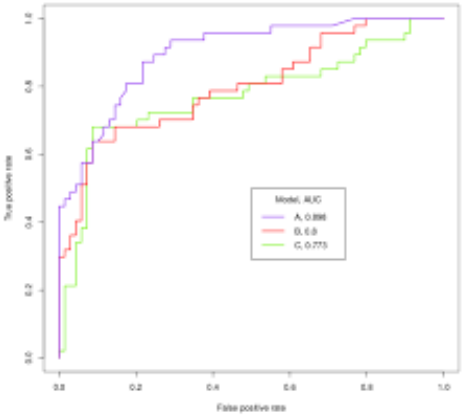
TP, FP
FN, TN

Accuracy: TP + TN / ALL
Especificity: TN / (TN + FP)
Sensivity = Recall: TP / (TP + FN)
Precision: TP / (TP + FP)
F1 (Precision + Recall): 2TP / (2TP + FP + FN)

Que es la curva ROC y AUC?

Receiver Operator Caracteristic, **Area under the curve**.
Es una métrica de medir la ratio entre True Positive vs. False Positive para clasificadores binarios.

Utilizando la curve se puede decidir donde poner el umbral que defina cuando un elemento entre en clase a o b, y cuanto sea la ratio entre TP y FP.



4. Clasificación con Naive Bayes

Que es el teorema de Bayes?

El teorema de Bayes se utiliza para relacionar la probabilidad de dos eventos A, B:

$$P(A\,|B)\ * P(B) = P(B\mid A) * P(A)$$

En caso de mas posibles eventos A_i se transforma en:
Con los A_i sean **independientes**, **mutualmente excluyentes y exlausivos**.

$$P(A_i|B) = \frac{P(B|A_i) * P(A_i)}{P(B)}$$

$$a\ posteriori = \frac{a\ priori *\ likelihood}{veosimilitud\ marginal}$$

Entonces se puede crear un clasificador que calcula dependiendo de los datos presentes/observaciones, que es la clase más probable.
Se llama **Naive** Bayes, porque se asume que los eventos Ai son independientes, y se puede calcular la probabilidad de aparecer solo dependiendo de B, no de los demás Ai.
Además, se asume que todos los variables/características son igual de importante.

Que es el estimador Laplace?

No se puede calcular la probabilidad en casos que no están en los datos de entrenamiento (división de cero) por eso se puede añadir un 'evento' virtual solo para ser capaz de acabar los cálculos.

5. Máquinas de vector de soporte

Que son las SVM?

Es un modelo de clasificación y regresión supervisado, desarrollado por Vladimir Vapnik 1963, y es el primer método que no se base en métodos bayesianas que dependen de la distribución de datos de entrenamiento.

La idea general es calcular un Hiperplano, una separación lineal de los datos.
El **Maximal-margin classifier** busca un hiperplano que maximiza la distancia entre los puntos de grupos distintos.

Softmargen: En la mayoría de los casos no hay una separación lineal, por eso se definen Hiperplanos que dejan una separación no perfecta, pagando un coste (parámetro C) para cada punto al lado equivocado del hiperplano.

Esa separación funciona mucho mejor en espacios de dimensión alta (por el mismo team de curse of dimensionality) y asi se transforma el input space -> **feature space** de alta dimensión. En teoría se puede encontrar una separación linear para cada datase si la dimensión es suficiente alta.

Kernel Trick: es la herramienta matemática para hacer esa transformación (eficaz y rápida)

6. Redes Neuronales

Que son redes neuronales y el teorema de la aproximación universal?

Son un método de aprendizaje automático, generando modelos de clasificación o regresión supervisados.

Son bioinspirados por redes neuronales de neo cortex, y se ha demostrado en el teorema de **la aproximación universal** que una red con una capa oculta puede aproximar cualquier función den función del numero de neuronas.

Una red es definida por

- **La topología**: #capas, #nodos, fully connected, feed forward, recurrent NN
- **La función de activación**: sigmoid, hyperbolic, linear
- **Algoritmo de entrenamiento**: Backpropagation (forward, backward pass) + gradient decent.

7. Árboles de decisión

Que es un árbol de decisión?

Un método de aprendizaje automático supervisado, útil por su transparencia. Basado en datos de entrenamiento, se separa el espacio de entrada por múltiples líneas rectas (cada Split).

No se puede calcular el árbol optimo (splits), por eso se utiliza un algoritmo **top-down greedy** como el 'recursive binary splitting', para encontrar la variable y el umbral donde hacer un Split.

Este recurisve binary splitting, es guiado por métricas como la **Entropia**, o **la ganancia de información** (incremento de información con un Split). Así se define si continuar hacer splits en hojas o no.

Tipos

- **Regression trees:** se calcula la media de todos valores de una hoja
- **Classification trees:** la clasa mayoría de la hoja

Los árboles tienen la tendencia de sobreajuste, y se puede compensar con **la poda del árbol**, guiada por el rendimiento en el test set. La desventaja del los árboles es que tienen una alta varianza, y se cambian mucho en función de los datos de entrenamiento.

DT vs. LR
Si los dataos tienen una estructura lineal, es mejor utilizar una regresión lineal.
El DT es mejor para problemas no lineales.

8. Métodos ensamble

Que es el bootstrapping?

Bootstrapping es un método estadístico, para estimar la variancia de una población en función de una única muestra. Se calcula la varianza de múltiples remuestros la muestra original como estimación de la varianza de la población.

Que son los métodos ensamble?

Son métodos en cual se utiliza un conjunto de modelos para reducir el sedo y la varianza.

- **Bagging**: (**B**ootstrap **Agg**regat**ing)** Se entrenan múltiples modelos clasificadores basado en partes del dataset (Bootstrapping) y se juntan las predicciones. Se puede calcular **el out-of-bag error** con los datos que no fuen utilizado para el entrenamiento. Se utiliza con modelos complejos para reducir su varianza.
- **Boosting:** Entrena una secuencia de modelo sencillos, con cada modelo nuevo, dando más peso a los puntos clasificados mal, en el modelo anterior. Ejemplo **Adaboost.** Se utiliza con modelos sencillos para reducir sesgo, mejorar la predicion.

9. Random Forests

Que son random forests?

Es un método bagging con árboles de decisión, muy reciente (2001).

Cada árbol del bagging, se construye teniendo en cuenta solo un subconjunto de los variables. En caso de regresión se utiliza la media, y en caso de clasificador la moda del ensamble para juntar la decisión de los diferentes modelos.

Se aplica el out-of-bag error rate, y como parámetro sirven #árboles y #de variables que tener en cuenta.

En comparación con los árboles de decisión, se incremente el sesgo y baja la varianza.

10. Agrupamiento

Que es el algoritmo de k-means?

Es un método de agrupamiento es un método **no supervisado** que intenta agrupar puntos de datos en grupos, con la meta de minimizar la diferencia entre elementos del mismo grupo, y maximizarlo entre grupos.

Se aplica a datos numéricos normalizados. Hay que especificar el numero de clústeres (k) a priori. Puede ser un valor que viene del contexto y un conocimiento previo.

Algoritmo es iterativo

0: Elija alazar k puntos centroides (solo una vez)
1: Asigna cada punto a su centroide más cercano.
2: Calcula el centroide para cada cluster basado en la media de sus puntos

Repita 1&2 hasta no se cambian las clusters

Que es a agrupamiento jerárquico?

Es otro método no supervisado que genera una orden jerárquica de los datos.
Aglomerativos: bottom-up
Dvisivos: top-down

Algoritmo Aglomerativo:

0: Cada punto empieza solo en su grupo propio
1: Junta dos grupos más cercanos

Repita 1 hasta que se llega a un grupo único. El resultado se puede visualizar como un árbol llamado **dendorgrama.**

11. Detección de anomalías

Como funciona la detección de anomalías?

Se puede ver como un método supervisado y no supervisado.
Basado en un conjunto de datos, se calcula **la función de densidad de probabilidad** y se calcula la probabilidad de aparecer como parte de esta distribución para cada punto. Utilizando **un umbral (épsilon – sensibilidad)** definido, se clasifican los puntos como normales y anomalías.

En comparación con aprendizaje supervisado, no hay suficientes datos de ambos grupos para entrenar un clasificador. Además, las anomalías puede ser muy distintos entre si.

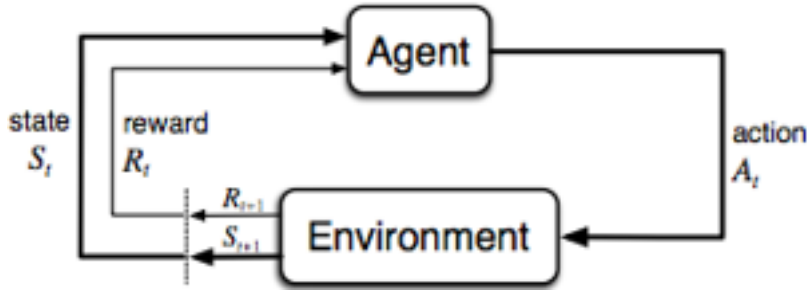
Como métrica se utiliza **F1 (Precision + Recall)**: 2TP / (2TP + FP + FN)

12. Aprendizaje por refuerzo

Que es el aprendizaje por refuerzo?

Es la tercera manera de aprendizaje automática y se distingue:
Aprendizaje supervisado: en AR importa una secuencia de acciones, en AS no.
Aprendizaje no supervisado: en AR hay una relación entre entrada y salida

Es un modelo donde un **Agente aprende del entorno** en función de recompensa.



Es modelado como un **Markov Decision Process (MDP)** en cual el estado siguiente y la recompensa solo depende del estado y accion actual. Se puede aplicar en entoernos observables o parcialmente ocultos.

La meta de AR es encontrar que acciones ejectuar en que estado para maximizar la recompensa.

Parametros:

- **Learning rate**: Entre 0-1 y especique cuanto se puede aprender de quada episodio.
- **Factor de descuento/Discount rate**: Entre 0-1, cuanto valen/importa una recompense futuras al recompensas de corto plazo.

Que es el Q-Learning?

Es una variación de aprendizaje por refuerzo muy popular con el tema de Deep learning, que básicamente esta generando una tabla, con todos los estados y posibles acciones.

Así el valor Q contiene la suma de todas las posibles recompensas futuras.

13. Parametrización automática

Que es la búsqueda de hiperparámetros y para que sirve?

Los hiperparámetros sirven controlar el over y underfitting, y optimizar los modelos.

Métodos

- **Optimización bayesiana:** generar una aproximación de la funcion (hiperparámetros->objectivo)
- **Optimización basada en gradiente**:
- **Optimización evolutiva**:

Las más comunes:

- **Grid search:** se puede paralelizar
- **Búsqueda aleatoria:** hay que tener en cuenta el criterio de parada

En general solo pocos parámetros importan de verdad.

La búsqueda aleatoria funciona muy bien, además para métodos con muchos parámetros! El grid search tiene el problema de 'curse of dimensionality' porque en el hiperespacio, los puntos de prueba están muy lejos de cada uno … y es espacio crece demasiado rápido.

Para modelos muy complejos la búsqueda aleatoria no funciona, y se están aplicando métodos como ***automatic sequential optimization,*** para generar un modelo del espacio de hiperparámetros, para guiar la búsqueda.