NOTAS

RECORDAR PREGUNTA DE ENTREVISTA DE TRABAJO

Algoritmo: Conjunto de instrucciones para realizar una atarea.

Modelo: Conjunto de algoritmos matemáticos que expresan relaciones y patrones entre variables y realizan predicciones.

El factor mas determinante de la calidad de un modelo es LOS DATOS. Entre un modelo perfecto y un modelo no tan bueno, la diferencia en poder de predicción es mínima, en los mas tops es de 99.6 vs 98%

Overfitting: El modelo simplemente se aprende todas las preguntas. Es cuando el modelo sabe demasiado las preguntas del examen pero no sabe responder nuevas preguntas. Es exceso de entrenamiento. Los errores en este son menos comunes pero cuando se equivoca, se equivoca por mucho.

Underfitting: El modelo necesita mas data. Nos equivocamos mucho y mal.

EL MEJOR MODELO ES EL QUE MEJOR FUNCIONA.

Tienes que probar varias veces, hay que aplicar el método científico. Hay que hacer cambios de uno a uno para ver si el funcionamiento mejora. A veces un detalle puede bajar o mejorar la performance. A veces podemos estarnos horas tuneando un modelo y no mejorará resultando en que el primero que hicimos es igual o mejor.

Cómo se entrena un modelo?

Se toma la data, el 80% se usa para entrenar el modelo y el 20% restante para el testeo. A veces puedo hacer 5 veces el mismo modelo con diferentes datos en el 80% y 20%.

Regularizacion: Es simplificar el modelo.

MÉTRICAS: No hay métricas mejor que otras. Todas son importantes: MSE, MAE, RMSE.

**AL HACER A PIE LA BÚSQUEDA DEL MINIMO ERROR**

Es el punto donde la derivada del error es 0. Si la derivada es negativa, hay que aumentar, si la derivada es positiva, hay que disminuir.

Pregunta: ¿No vendría perfecto aquí un while que dependa de que las derivadas se hagan cero?

Learning rate base (recomendado): 0.001

¿Hay alguna forma parametrizada de comprobar que hay muchos errores groseros?

**LA MALDICIÓN DE LA MULTIDIMENSIONALIDAD**

Teoría que sostiene que mas columnas tienen un impacto negativo si no aportan información nueva.

**Machine Learning – Regresión Lineal**

Familia de modelos de ML supervisados donde la tarea consiste en predecir un número.

**Machine Learning - Clasificación**

Familia de modelos de ML supervisados donde la tarea consiste en predecir una categoría.

**Machine Learning – Clasificación**

**Métricas:**

**La mas importante:**

Accuracy = correct predictions / all predictions – No llamar precisión

**Nota**: Accuracy no es bueno cuando el dataset esta desbalanceado, es decir, hay muchos datos de una categoría y pocos de otra.

Matriz de confusión: Muy útil para ver todo.

Precision = Es la accuracy de una categoría.

USO DEL HEATMAP

Vemos correlaciones fuertes entre las features. En este caso por ejemplo, S1 y S2 tienen una correlación fuerte, por ello nos quedamos como la que tenga la correlación con la Y mas alta, en este caso, S1.

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

¿Qué es estandarizar?

Conjunto de métodos y técnicas para mejorar la generalización del modelo

¿Qué es generalizar?

Hacer que el modelo funcione mejor ante preguntas que no ha visto. Ponemos preguntas de examen (test) nuevas. Si solo se memorizó cual era cual, esta mal.

¿Cómo sabemos si lo estamos haciendo bien?

Con el test. El error de train no nos dice todo. Recordar ejemplo de modelos distintos. Un modelo puede tener menos error en train, pero ser mucho peor en test, que es el que importa. Esto a causa del overfitting, que es cuando el modelo se aprende las respuestas, no aprende a generalizar. Hay que aprender a distinguir memoria de aprendizaje.

IDEA

Al generalizar/regularizar. Conviene tener “m” no tan grandes porque cualquier cambio en la variable dispara el crecimiento del valor. No PODEMOS DEPENDER TANTO DE UNA SOLA VARIABLE. Recordar ejemplo de marca y vehículos. Si dependemos mucho de la variable marca, cuando venga un modelo de coche nuevo, lo haremos mal.

L2: Los gordos bajaran.

L1: Todos bajarán. Los pequeñitos se van a cero. NOTA: Cuando probemos, los valores extremos pueden crecer en vez de decrecer, es decir, no necesariamente van a bajar al generalizar.

Ridge (L2) y Lasso (L1) Son las implementaciones de los métodos (L1 y L2).

Np.linspace: Te parte la distancia entre 2 numeros en un número de tramos.

Logspace: Es como elevar un número al valor. Elevar a la 3 por ejemplo es 1000.

Al hacer crecer alfa, llega a un punto donde los “m” se hacen tan insignificante y cercanos a cero que el error deja de subir.

* [ ] Regularización: es un conjunto de técnicas que están hechas para que el modelo generalice mejor.
* [ ] Generalizar: Mejorar el modelo para que sea capaz de hacer predicciones con datos que no estaban en el entrenamiento