Ciencia de Datos

Temario del plan de estudios

Objetivos de Aprendizaje

Se espera que al finalizar el cursado del módulo los estudiantes sean capaces de:|

Diseñar, desarrollar e implementar técnicas de modelos predictivos y reconocimiento de segmentos, entre otros.|Diseñar visualizaciones de informaciones acertadas y correctamente realizadas.|Conocer los diferentes tipos de datos existentes así como el tipo de análisis correspondiente.|Manejar las técnicas de clasificación y visualización de datos.

Contenidos|Bloque La Ciencia de Datos|Definición y conceptos de ciencia de datos. Problemáticas específicas vinculadas al uso y manejo de la información. Características y procesos propios de las organizaciones. Modelos tradicionales de gestión de la información en las empresas y/u organizaciones. Cultura analítica organizacional basada en la ciencia de datos. Ciclo de vida del dato (captura, preprocesamiento, análisis y visualización). Preparación de los datos. Validación y evaluación de resultados. Extracción y selección de atributos. Protocolos de validación. Calidad, privacidad y seguridad de los datos. Ética en ciencia de datos. Ciencia de datos como factor clave para la autonomía tecnológica, desarrollo económico y competitividad en las industrias.|Bloque Metodología para análisis

Uso actual de los tableros de control: ventajas y desventajas. La Ciencia de Datos como herramienta de análisis predictivo para la optimización de proyectos y/o negocios.|Diferencias entre Inteligencia de Negocios y Análisis Predictivo. Capacidad analítica para el manejo de la información en la gestión de negocios La visualización y transformación de la información como base innovadora para la toma de decisiones. La representación visual de datos como variable de ahorro de tiempo en las organizaciones.

Ciencia de datos

¿Qué es la ciencia de datos?

La ciencia de datos es el estudio de datos con el fin de extraer información significativa para empresas. Es un enfoque multidisciplinario que combina principios y prácticas del campo de las matemáticas, la estadística, la inteligencia artificial y la ingeniería de computación para analizar grandes cantidades de datos.

¿Por qué es importante la ciencia de datos?

Las empresas modernas están inundadas de datos; hay una proliferación de dispositivos que pueden recopilar y almacenar información de manera automática. Los sistemas on line y los portales de pago capturan más datos en los campos del comercio electrónico, la medicina, las finanzas y cualquier otro aspecto de la vida humana. Disponemos de grandes cantidades de datos de texto, audio, video e imágenes.

¿Para qué se utiliza la ciencia de datos?

La ciencia de datos se utiliza para estudiar los datos de cuatro maneras principales Análisis descriptivo, Análisis de diagnóstico, Análisis predictivo, Análisis prescriptivo

Análisis descriptivo

El análisis descriptivo examina los datos para obtener información sobre lo que ha ocurrido u ocurre en el entorno de datos. Se caracteriza por las visualizaciones de datos, como los gráficos circulares, de barras o líneas, las tablas o las narraciones generadas. Por ejemplo, un servicio de reserva de vuelos registra datos como el número de billetes reservados cada día. El análisis descriptivo revelará los picos y las caídas de las reservas, así como los meses de alto rendimiento del servicio.

Análisis de diagnóstico

El análisis de diagnóstico es un examen profundo y detallado de datos para entender por qué ha ocurrido algo. Se caracteriza por técnicas como el análisis detallado, el descubrimiento y la minería de datos o las correlaciones. Se pueden llevar a cabo varias operaciones y transformaciones de datos en un determinado conjunto con el fin de descubrir patrones únicos en cada una de estas técnicas. Por ejemplo, el servicio de vuelos podría hacer el análisis detallado de un mes con un rendimiento particularmente alto para entender

mejor el pico de reservas. Esto puede revelar que muchos clientes visitan una determinada ciudad para asistir a un evento deportivo mensual.

Análisis predictivo

El análisis predictivo utiliza los datos históricos para hacer previsiones precisas sobre los patrones de datos que pueden producirse en el futuro. Se caracteriza por técnicas como el machine learning, la coincidencia de patrones y el modelado predictivo. Por ejemplo, el equipo de servicios de vuelo podría utilizar la ciencia de datos para predecir los patrones de reserva de vuelos del año siguiente al inicio de cada año. El programa o algoritmo de la computadora pueden examinar datos anteriores y predecir picos de reservas de determinados destinos en mayo. Al anticiparse a las futuras necesidades de viaje de los clientes, la empresa podría empezar desde febrero a hacer publicidad específica para esas ciudades.

Análisis prescriptivo

El análisis prescriptivo no solo predice lo que es probable que ocurra, sino que sugiere una respuesta óptima para ese resultado. Puede analizar las posibles implicaciones de las diferentes alternativas y recomendar el mejor curso de acción.

De vuelta al ejemplo de la reserva de vuelos, el análisis prescriptivo podría examinar las campañas de marketing históricas para maximizar la ventaja del próximo pico de reservas. Un científico de datos podría proyectar los resultados de las reservas de diferentes niveles de gasto en varios canales de marketing. Estas previsiones de datos dan a la empresa de reserva de vuelos una mayor confianza en sus decisiones de marketing.

¿Cuáles son los beneficios de la ciencia de datos para las empresas?

La ciencia de datos revoluciona el funcionamiento de las empresas. Muchas empresas, independientemente de su tamaño, necesitan una sólida estrategia de ciencia de datos para impulsar el crecimiento y mantener una ventaja competitiva. Algunos beneficios clave son

descubrir patrones desconocidos de transformación, innovar con nuevos productos y soluciones, Optimización en tiempo real

La ciencia de datos revoluciona el funcionamiento de las empresas. Muchas empresas, independientemente de su tamaño, necesitan una sólida estrategia de ciencia de datos para impulsar el crecimiento y mantener una ventaja competitiva. Algunos beneficios clave son:

Descubrir patrones desconocidos de transformación:|La ciencia de datos permite a las empresas descubrir nuevos patrones y relaciones con el potencial de transformar la organización. Puede revelar cambios de bajo coste en la administración de recursos para obtener el máximo impacto en los márgenes de beneficio. Por ejemplo, una empresa de comercio electrónico utiliza la ciencia de datos para descubrir que se generan demasiadas consultas de clientes fuera del horario comercial. Las investigaciones revelan que es más probable que los clientes compren si reciben una respuesta rápida en lugar de una respuesta al día siguiente. Al implementar un servicio de atención al cliente las 24 horas del día, los 7 días de la semana, la empresa aumenta sus ingresos en un 30 %.

Innovar con nuevos productos y soluciones: La ciencia de datos puede revelar problemas que de otro modo pasarían desapercibidos. Mejor información sobre las decisiones de compra, los comentarios de los clientes y los procesos empresariales puede impulsar la innovación en las operaciones internas y las soluciones externas.

Por ejemplo, una solución de pago on line utiliza la ciencia de datos para analizar los comentarios que hacen los clientes sobre la empresa en redes sociales. Los análisis revelan que los clientes olvidan las contraseñas durante los periodos de pico de compra y que no están satisfechos con el actual sistema de recuperación de contraseñas. La empresa puede innovar para obtener una mejor solución y ver un aumento significativo en la satisfacción del cliente.

Optimización en tiempo real: Para las empresas es un gran reto responder en tiempo real a las condiciones cambiantes. Esto puede causar importantes pérdidas o interrupciones en la actividad empresarial. Por ejemplo, una compañía de envíos que usa camiones utiliza la ciencia de datos para reducir el tiempo de inactividad si los camiones se rompen. Identifican las rutas y los patrones de turnos que propician averías más rápidas y ajustan los horarios de los camiones. Además, crean un inventario de piezas de repuesto comunes que se necesitan sustituir con frecuencia para que los camiones se puedan reparar con mayor rapidez.

El proceso de ciencia de datos

Paso 1: encuadre el problema.
Paso 2: recopile los datos sin procesar necesarios para su problema.
Paso 3: Procesar los datos para su análisis.
Paso 4: Explora los datos.
Paso 5: realizar un análisis en profundidad.
Paso 6: Comunicar los resultados del análisis.

Lo primero que debemos hacer antes de resolver un problema, es definir exactamente de qué se trata. Debe poder traducir las preguntas sobre datos en algo procesable.

Supongamos que está resolviendo un problema para el vicepresidente de ventas de su empresa. Debe comenzar por comprender sus objetivos y el por qué subyacente detrás de sus preguntas sobre datos. Luego deberías averiguar cómo es el proceso de ventas y quiénes son los clientes.

Deberías hacer preguntas como las siguientes:

¿Quiénes son los clientes?, ¿Por qué compran nuestro producto?, ¿Cómo predecimos si un cliente va a comprar nuestro producto?, ¿En qué se diferencian los segmentos que se están desempeñando bien y aquellos que se están desempeñando por debajo de las expectativas?,

¿Cuánto dinero perderemos si no vendemos activamente el producto a estos grupos?

En las respuestas, el vicepresidente de ventas podría revelar que quiere comprender por qué ciertos segmentos de clientes han comprado menos de lo esperado. Su objetivo final podría ser determinar si continuar invirtiendo en estos segmentos o quitarles prioridad. Querrá adaptar su análisis a ese problema y descubrir ideas que puedan respaldar cualquiera de las conclusiones.

Es importante que al final de esta etapa tengas toda la información y el contexto que necesitas para resolver este problema.

Paso 2: recopile los datos sin procesar necesarios para su problema

Una vez que haya definido el problema, necesitará datos que le brinden la información necesaria para solucionar el problema con una solución. Esta parte del proceso implica pensar qué datos necesitará y encontrar formas de obtenerlos, ya sea consultando bases de datos internas o comprando conjuntos de datos externos.

Es posible que descubra que su empresa almacena todos sus datos de ventas en un CRM o en una plataforma de software de gestión de relaciones con los clientes. Puede exportar los datos del CRM en un archivo CSV para su posterior análisis.

Paso 3: Procesar los datos para su análisis.

A menudo, los datos pueden ser bastante confusos valores establecidos en nulo aunque en realidad son cero, valores duplicados y valores faltantes. Querrá resolver los siguientes errores: Valores faltantes, quizás clientes sin una fecha de contacto inicial, valores corruptos, como entradas no válidas, quizás la base de datos no tenga en cuenta las diferentes zonas horarias de los usuarios o tal vez tenga fechas que no tengan sentido, como datos registrados antes de que comenzaran las ventas.

Si se detecta algo que no tiene sentido, se deberán eliminar esos datos o reemplazarlos con un valor predeterminado. Necesitaremos usar la intuición: si un cliente no tiene una fecha de contacto inicial, ¿tiene sentido decir que NO hubo fecha de contacto inicial? ¿O hay que buscar al vicepresidente de ventas y preguntarle si alguien tiene datos sobre las fechas de contacto iniciales que faltan del cliente?

Una vez que haya terminado de trabajar con esas preguntas y limpiar sus datos, estará listo para el análisis de datos exploratorios (EDA).

Paso 4: Explora los datos

En esta etapa se buscan patrones que pueden ayudar a explicar por qué se reducen las ventas para este grupo (no suelen ser muy activos en las redes sociales, y pocos de ellos tienen cuentas de Twitter o Facebook, también podríamos ver que la mayoría de ellos son mayores a la media , etc.

A partir de ésos comportamientos podríamos rastrear patrones que podríamos analizar más profundamente.

Paso 5: realizar un análisis en profundidad

En este paso del proceso es donde tendrá que aplicar sus conocimientos estadísticos, matemáticos y tecnológicos y aprovechar todas las herramientas de ciencia de datos a su disposición para analizar los datos y encontrar toda la información que pueda.

En este caso, es posible que deba crear un modelo predictivo que compare su grupo de bajo rendimiento con su cliente promedio. Es posible que descubra que la edad y la actividad en las redes sociales son factores importantes para predecir quién comprará el producto.

Podríamos darnos cuenta de que la empresa ha concentrado su comunicación en redes sociales, con mensajes dirigidos a audiencias más jóvenes. Un cambio de estrategia de marketing (un contacto telefónico) podría cambiar todo para mejor. Esto es algo que deberá informar a su vicepresidente de ventas.

Paso 6: Comunicar los resultados del análisis.

La etapa de **comunicación de resultados** en el procesamiento de datos es crucial, ya que permite traducir los hallazgos técnicos en información clara, útil y accionable para los interesados (stakeholders). Es el puente entre el análisis de datos y la toma de decisiones basada en dichos análisis

Aquí se elabora una historia que vincule los datos con su respaldo. Se empieza explicando las razones del bajo rendimiento del grupo demográfico de mayor edad. Esto se relaciona con las respuestas que le dio su vicepresidente de ventas y los conocimientos que descubrió a partir de los datos. Luego se pasa a soluciones concretas que aborden el problema: podríamos trasladar algunos recursos de las redes sociales a llamadas personales. Lo unes todo en una narrativa que resuelve el dolor de tu vicepresidenta de ventas: ahora tiene claridad sobre cómo puede recuperar las ventas y alcanzar sus objetivos.

¿A qué retos se enfrentan los científicos de datos?

Varios orígenes de datos: Los diferentes tipos de aplicaciones y herramientas generan datos en varios formatos. Los científicos tienen que limpiar y preparar los datos para que sean coherentes. Esto puede resultar tedioso y llevar mucho tiempo.

Entender el problema de la empresa

Los científicos de datos tienen que trabajar con varias partes interesadas y con administradores empresariales para definir el problema que se debe resolver. Esto puede

suponer un reto, particularmente en empresas grandes que cuentan con múltiples equipos de trabajo con necesidades diferentes.

Eliminación del sesgo: Las herramientas de machine learning no son completamente precisas, por lo que puede existir cierta incertidumbre o sesgo. Los sesgos son desajustes en el comportamiento de las predicciones o los datos de entrenamiento del modelo entre diferentes grupos, como la edad o el nivel de ingresos. Por ejemplo, si una herramienta se entrena principalmente con datos de personas de mediana edad, puede ser menos preciso cuando se hagan predicciones que impliquen a personas más jóvenes o mayores. El ámbito del machine learning ofrece la oportunidad de abordar los sesgos detectados y midiéndose en los datos y el modelo.

Estadística Descriptiva

Estadística

La Estadística es la parte de las Matemáticas que se encarga del estudio de una determinada característica en una población, recogiendo los datos, organizándose en tablas, representarlos gráficamente y utilizándolos para sacar conclusiones de dicha población.

El término estadística también se usa para denotar los datos o los números que se obtienen de esos datos; por ejemplo, los promedios. Así, se habla de estadísticas de empleo, estadísticas de accidentes, etc. La ciencia de datos utiliza la estadística para extraer conocimiento de los datos. En esta unidad veremos algunos conceptos y técnicas de estadística comunes.

Población vs Muestra

Cuando se recolectan datos sobre las características de un grupo de individuos o de objetos, puede ser imposible observar todo el grupo, en especial si se trata de un grupo grande. En vez de examinar todo el grupo, al que se le conoce como población o universo, se examina sólo una pequeña parte del grupo, al que se le llama muestra. Las poblaciones pueden ser finitas o infinitas.

Característica de la Muestra

Una Muestra para que sea válida tiene dos características aleatoriedad que es una muestra aleatoria cuando cada miembro de la muestra se elige de entre la población estrictamente al azar. Representatividad es una muestra representativa cuando es un subconjunto de la población que refleja fielmente a los miembros de toda la población.

Estadística Descriptiva o Deductiva vs Estadística Inductiva o Inferencial

A la parte de la estadística que únicamente trata de describir y analizar un grupo dado, sin sacar ninguna conclusión ni hacer inferencia alguna acerca de un grupo más grande, se le conoce como estadística descriptiva o deductiva.

Si, en cambio, a partir de una muestra, logramos inferir conclusiones que son válidas para toda la población, estamos aplicando estadística inductiva o inferencial.

Variables cuantitativas y cualitativas: discretas y continuas. Dominio

Una variable cualitativa o categórica describe cualidades, circunstancias o características de un objeto o persona. No pueden ser medidas en números y se pueden distinguir dos tipos:

Variable cualitativa nominal: no admiten un criterio de orden (cuatro estaciones: invierno, primavera, verano, otoño).

Variable cualitativa ordinal: tiene una modalidad no numérica, pero existe un orden (calificación conceptual – desaprobado, aprobado, sobresaliente)

Una variable cuantitativa o numérica puede tomar una serie de valores y admiten operaciones aritméticas. Puede ser variable discreta o continua.

Una **variable discreta** es un tipo de variable que no puede tomar algunos valores, es decir, una variable discreta solo puede tomar un número finito de valores entre dos valores cualesquiera. Por ejemplo, el número de personas en una habitación es una variable discreta porque no puede haber 3,92 personas, la variable solo puede ser un número entero.

Una variable que puede tomar cualquier valor entre dos números cualquiera es una variable

continua (por ejemplo: longitud o temperatura); de lo contrario es una variable discreta (como cantidad de alumnos o unidades vendidas).

A los valores válidos que puede tomar una variable, se le denomina dominio de esa variable. Por ejemplo: La cantidad de agua que entra en un recipiente es una variable cuantitativa continua, cuyo dominio es desde 0, hasta la capacidad total del recipiente.

Países de Europa es una variable cualitativa nominal, pero aún así tiene un dominio: España, Portugal, Francia, Italia, Alemania, etc.

Instrumentos o medidas de la Estadística Descriptiva

Las principales medidas de la estadística descriptiva son las siguientes:

• Razones, tasas y porcentajes: son medidas relativas que condensan información sobre la incidencia de una característica entre un grupo de unidades.

• Distribución de frecuencia: forma de agrupación de los datos, en la cual estos se presentan en clases y cada clase exhibe su respectiva frecuencia.

• Medidas de posición o de tendencia central: se dividen en promedios matemáticos el aritmético, geométrico y armónico. Promedios no matemáticos: la mediana y la moda.

• Medidas de dispersión o variabilidad: para las variables cuantitativas, las medidas de dispersión que se pueden identificar son la desviación media, la desviación estándar o desviación típica, los rangos intercuartílicos y los valores mínimos y máximos.

Medidas de la Estadística Descriptiva - Razones, Tasas y Porcentajes

Razón

Se define la razón como el valor que indica la relación cuantitativa entre dos cantidades. Por ejemplo, si en una zona geográfica determinada existen 40.000 niños escolarizados y 10.000 no escolarizados, la razón de escolarizado y no escolarizado vendría expresada por

el cociente: 40000/10000 = 4, De acuerdo con el resultado, se diría entonces que por cada cuatro niños escolarizados, hay un niño no escolarizado.

Tasa

En la proporción o tasa, a diferencia del índice anterior, el denominador del cociente es el número total de unidades enunciadas. Tomando el ejemplo anterior para la razón, las proporciones de niños escolarizados y no escolarizados serían: Niños escolarizados = 40000 / 50000 = 0.8 | Niños no escolarizados = 10.000 / 50.000 = 0.2

Debe observarse que al sumar las dos tasas obtenidas (0,80 + 0,20), el resultado es uno (1), ya que son proporciones complementarias.

Porcentaje

Como se observa en el ejemplo de la tasa, la solución viene expresada en valores decimales, y si bien desde el punto de vista estadístico no es un inconveniente, usualmente los resultados se presentan en porcentajes. Es por ello que se acostumbra a multiplicar las proporciones por 100, para convertir los valores decimales en porcentajes. Niños escolarizados = 0.8 * 100% = 80 % | Niños no escolarizados = 0.2 * 100% = 20%

Distribución

La distribución nos muestra la frecuencia de diferentes resultados (o puntos de datos) en una población o muestra. Podemos mostrarla como números en una lista o tabla, o podemos representarla gráficamente. Como ejemplo básico, la siguiente lista muestra el número de personas con diferentes colores de cabello en un conjunto de datos de 286 personas.

Cabello castaño: 130

Cabello negro: 39

Cabello rubio: 91

Cabello pelirrojo: 13

Canoso: 13

También podemos representar esta información visualmente, por ejemplo, en un gráfico circular.

El uso de visualizaciones es una práctica común en estadística descriptiva. Nos ayuda a detectar patrones o tendencias más fácilmente en un conjunto de datos.

Distribución de frecuencia

• Datos en bruto: son los datos recolectados que aún no se han organizado. Por ejemplo, las edades de 100 miembros practicantes de tai-chi para adultos mayores.

• Ordenación: se le llama a los datos numéricos en bruto, dispuestos en orden creciente o decreciente.

La diferencia entre el número mayor y el número menor, se le llama rango de los datos. Por ejemplo, si la mayor edad de entre los 100 miembros es de 74 años, y la menor de 60 años, el rango es 74 – 60 = 14 años.

Al organizar gran cantidad de datos, se suele distribuir en clases o categorías. Llamamos frecuencia de clase a la cantidad de datos que pertenecen a cada clase.

Si presentamos esta información en forma de tabla, obtenemos una distribución de frecuencias o tabla de frecuencias.

Histogramas y polígonos de frecuencias

Los histogramas y los polígonos de frecuencias son dos maneras de representar gráficamente las distribuciones de frecuencias.

1. Un histograma o histograma de frecuencias consiste en un conjunto de rectángulos que tienen:

a) sus bases sobre un eje horizontal (el eje X), con sus centros coincidiendo con las marcas de clase de longitudes iguales a la amplitud del intervalo de clase, y

b) áreas proporcionales a las frecuencias de clase.

Histogramas y polígonos de frecuencias

Un polígono de frecuencias es una gráfica de línea que presenta las frecuencias de clase graficadas contra las marcas de clase. Se puede obtener conectando los puntos medios de las partes superiores de los rectángulos de un histograma.

Medidas de posición o de tendencia central

Tendencia central es el nombre de las mediciones que miran los valores centrales típicos dentro de un conjunto de datos. Esto no solo se refiere al valor central dentro de un conjunto de datos completo, que se denomina mediana, sino que es un término general utilizado para describir una variedad de medidas centrales. Por ejemplo, podría incluir mediciones centrales de diferentes cuartiles de un conjunto de datos más grande.

• La moda: el valor que aparece con más frecuencia en el conjunto de datos.
• La mediana: el valor central o medio del conjunto de datos.
• La media: el valor promedio de todos los puntos de datos.

La Media

Es el valor promedio de un conjunto de datos numéricos.
media x = (x1 +x2+ x3+ …+xn)/N

La media aritmética para la siguiente serie de cifras se determinaría de esta manera:
Cifras: 5, 9, 10, 12, 16, 19, 22, 27.

N = 8 (el número datos).

La mediana

La mediana se define como el valor que divide una distribución de manera que un número igual de términos quede a cada lado.

La mediana de un conjunto de números ordenados es el valor central o la media de los dos valores centrales.

Buscamos la mediana en dos series de conjuntos A y B:

A = [56, 5, 88, 2, -10, 73, 11, 8, 33]
B = [22, 9, 5, 32, 14, -3, 1, 12, 2, 9]

El primer paso es ordenar los conjuntos:

Aord = [-10, 2, 5, 8, 11, 33, 56, 73, 88]
Bord = [-3, 1, 2, 5, 9, 7, 12, 14, 22, 32]

Busco el elemento central de Aord porque tiene una cantidad de elementos impar.

Aord = [-10, 2, 5, 8, 11, 33, 56, 73, 88]

Entonces el 11 es la mediana del conjunto Aord, y determina una partición en el conjunto con la misma cantidad de elementos a izquierda y derecha.

En el conjunto Bord tenemos una cantidad par de elementos, por lo que debemos hallar los dos centrales, y calcular su media:

Bord = [-3, 1, 2, 5, 7, 9, 12, 14, 22, 32]

media(7,9) = (7+9)/2 = 8

Decimos entonces que la mediana del conjunto Bord es 8.

Moda

Se define como el valor de la serie que más se repite, el valor más típico. Es decir que la moda es el punto donde la concentración es máxima.

En un conjunto de datos puede no haber moda, y si la hay, puede que no sea única.

La moda del conjunto [2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18] es 9. Tiene sólo una, por lo que se llama unimodal.

El conjunto [3, 5, 8, 10, 12, 15, 16] no tiene moda.
El conjunto [2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 9] tiene dos modas, 4 y 7, por lo que se le llama bimodal.

Significado Media – Mediana y Moda

Estas tres medidas nos indican la forma en que están distribuidos nuestros datos

Distribución normal, gaussiana, de Laplace-Gauss o distribución de Gauss. Coinciden media, mediana y moda. La curva es acampanada y simétrica respecto de estos parámetros.

Rango - Valores mínimos y máximos

Los valores mínimos y máximos son los valores más altos y más bajos en un conjunto de datos o cuartil. El rango es el intervalo entre el valor máximo y el valor mínimo.

En el ejemplo, tenemos que:

Valor mínimo: 2
Valor máximo: 10
Rango: 8

El rango no proporciona suficiente información sobre la variabilidad de los datos y puede ser una medida Engañosa.

Rango Intercuartil, deciles y percentiles

Quartiles

En un conjunto de datos en el que éstos se hallan ordenados de acuerdo con su magnitud, el valor del medio (o la media aritmética de los dos valores del medio), que divide al conjunto en dos partes iguales, es la mediana. Continuando con esta idea se puede pensar en aquellos valores que dividen al conjunto de datos en cuatro partes iguales. Estos valores, denotados Q1, Q2 y Q3 son el primero, segundo y tercer cuartiles, respectivamente; el valor Q2 coincide con la mediana.

Percentiles

Son los valores que dividen a un conjunto de datos ordenados en cien partes iguales. De manera que un percentil indica el valor por debajo del cual se encuentra un porcentaje del conjunto de datos.

De igual manera, los valores que dividen al conjunto en diez partes iguales son los deciles y se denotan D1, D2, . . . , D9, y los valores que dividen al conjunto en 100 partes iguales son los percentiles y se les denota P1, P2, . . . , P99. El quinto decil y el percentil 50 coinciden con la mediana. Los percentiles 25 y 75 coinciden con el primero y tercer cuartiles, respectivamente.

A los cuartiles, deciles, percentiles y otros valores obtenidos dividiendo al conjunto de datos en partes iguales se les llama en conjunto cuantiles.

Los cuartiles

Los cuartiles son los tres elementos de un conjunto de datos ordenados que dividen el conjunto en cuatro partes iguales.

Características de los cuartiles:

• El cuartil 1 (Q1) es el percentil 25 (P25). El 25 % de los datos son menores o iguales a Q1.
• El cuartil 2 (Q2) es la mediana y el percentil 50 (P50). El 50 % de los datos son menores o iguales a Q2.
• El cuartil 3 (Q3) es el percentil 75 (P75). El 75 % de los datos son menores o iguales a Q3.

Rango intercuartil o IQR

Es la diferencia entre el tercer y primer cuartil (r = Q3 - Q1). El IQR se puede utilizar para encontrar valores atípicos (valores en el conjunto que se encuentran significativamente fuera del valor esperado). Los valores que se encuentran a más de 1,5 veces el IQR de cualquiera de los extremos del IQR (Q1 o Q3) se consideran valores atípicos, como se muestra en la siguiente figura:

Por lo tanto, el rango de valores esperado es:

[Q1 – 1.5 (IQR), Q3 + 1.5 (IQR)]

Todo lo que esté fuera del rango de valores anterior es un valor atípico.

Ejemplo: Busque valores atípicos para el siguiente conjunto de datos:

{1, 3, 4, 6, 13, 20, 25, 26, 28, 62, 95}
Q2 = 20 (es la mediana y, dado que hay 11 elementos en el conjunto, Q2 es el valor medio)
Q1 = 4 (es la mediana del primer 25% de los valores (primer elemento hasta Q1)
Q3 = 28 (es la mediana del último 25% de los valores (Q3 hasta el elemento final).

El primer trimestre está resaltado en verde, el segundo trimestre en rojo y el tercer trimestre en azul: {1, 3, 4 , 6, 13, 20 , 25, 26, 28 , 62, 95}

IQR es la diferencia entre Q3 y Q1: IQR=28 – 4=24

Por lo tanto, cualquier valor fuera del rango

[4 – 1,5 (24), 28 + 1,5 (24)] = [-32, 64]

es un valor atípico. Todos los datos, excepto el 95, se encuentran dentro del rango anterior. Por lo tanto, 95 es el único valor atípico del conjunto.

Desviación Estándar

Esto nos muestra la cantidad de variación o dispersión. Una desviación estándar baja implica que la mayoría de los valores están cerca de la media. La desviación estándar alta sugiere que los valores están más dispersos. Dado un dataset = {x1, x2,... xn }, definimos la media (también llamada μ). Conceptualmente, lo que se esta haciendo es calcular la distancia entre la media y un punto cualquiera del set de datos (xi). La fórmula es:

poblacion:$\sigma = \sqrt{\Sigma (x_i - \mu)^2 / N}$|muestra:$s = \sqrt{\Sigma (x_i - \bar{x})^2 / (n - 1)}$

→ cuadrado de la distancia entre la media y un punto xi

Varianza

La desviación estándar viene dada por un concepto llamado varianza , que es la sumatoria de los cuadrados de las distancias entre todos los puntos y la media, dividido por la cantidad de elementos.

La desviación estándar no es más que la raíz cuadrada de la varianza. La varianza es una medida de dispersión que se expresa en las unidades de la variable al cuadrado, hecho que imposibilita su visualización en una gráfica como el histograma. Es por eso que en muchas ocasiones es mejor utilizar como medida de dispersión a la desviación estándar , calculada como la raíz cuadrada de la varianza y que, en consecuencia, se expresa en las mismas unidades que la variable.

Significado en distribuciones asimétricas o sesgadas

En este caso no podemos aplicar la fórmula del método de detección de outliers con el rango intercuartil. Lo que suele hacerse es generalizar la noción de cómo definir el criterio a partir del cual los datos se consideran anómalos por medio de una función que nos permita calcular perfectamente el sesgo de la distribución. En conclusión: cuando tenemos una distribución simétrica se suele trabajar con la desviación estándar, en cambio, con

distribuciones sesgadas, es más conveniente utilizar el rango intercuartil para determinar los datos válidos

Asimetría o Sesgo

Es una medida de la simetría de un conjunto de datos. El sesgo estadístico es la diferencia que se produce entre un estimador matemático y su valor numérico, una vez realizado un análisis. Si tuvieras que trazar una curva de campana y la cola de la derecha fuera más larga y más gruesa, la llamaríamos sesgo positivo. Si la cola de la izquierda es más larga y más gruesa, la llamaríamos sesgo negativo.

Relación entre el Sesgo y las Medidas de Tendencia Central

La media es la medida de tendencia central más sensible al sesgo. Si agregáramos datos con valores muy altos (o muy bajos) de la variable, estos tendrían un fuerte impacto en la sumatoria de la ecuación de la media y, en consecuencia, en el resultado final. Datos con valores muy altos incrementarían el valor de la media y en el caso opuesto la media disminuiría en comparación con el caso no sesgado. Por lo general, en una distribución con sesgo positivo (o a la derecha) la media tendrá un valor mayor que la mediana, y la mediana tendrá un valor mayor que la moda. En una distribución con sesgo negativo la media tendrá un valor menor que la mediana, y la mediana tendrá un valor menor que la moda.

Curtosis (ck)

Es la medida que define qué tan pronunciada es la punta (o pico) en una distribución. Desde el punto de vista de la Ciencia de Datos, mide si las colas de una distribución dada contienen valores extremos (también conocidos como valores atípicos). Si una cola carece de valores atípicos, podemos decir que tiene una baja curtosis. Si un conjunto de datos tiene muchos valores atípicos, podemos decir que tiene una curtosis alta.

Para una distribución dada, la curva normal sirve como patrón de referencia.

• Una distribución con ck = 0 (distribución normal) se le denomina mesocúrtica.
• Una distribución en donde el ck > 0 se le denomina leptocúrtica, lo que implica que es más puntiaguda y con colas más anchas que la distribución normal de referencia.
• Una distribución en donde el ck < 0 , platicúrtica, es más aplanada y con colas menos anchas que las de la distribución normal.

Correlación entre variables

Eventualmente vamos a querer conocer si existe una relación entre dos o más variables de un dataset, entendiendo por relación a una “variación conjunta”. Si este es el caso, podríamos ver que si una de las variables aumenta o disminuye su valor, que la otra también lo hace.

La **covarianza** es una medida que intenta cuantificar esa relación.

Es posible que x e y estén utilizando diferentes magnitudes o unidades, por lo que debemos buscar una forma más estándar para normalizar el grado de variación de cada variable. Esto nos introduce al llamado coeficiente de correlación ρ, siendo la covarianza dividida la desviación standard de x, multiplicada por la desviación standard de y:

Este coeficiente se denomina correlación lineal o de Pearson, y es una cantidad adimensional. Por lo tanto, si cambiamos la unidad de x e y, el valor del coeficiente seguirá siendo el mismo.

Podemos ver algunos de los distintos tipos de correlaciones que podemos encontrar, variando desde r=-1 (correlación negativa perfecta), hasta r=1 (correlación positiva perfecta).

Justo en medio podemos encontrar r=0 (no hay ningún tipo de correlación), y diversos valores -1<r<0 y 0<r<1, que indican diferentes grados de correlación intermedios: altas negativas, bajas negativas, bajas positivas y altas positivas.

Esta información nos sirve para realizar reducción de datos: si dos variables están fuertemente correlacionadas, es posible que ambas estén aportando la misma información. Deberíamos eliminar una de ellas de nuestro análisis.

Conclusiones

• La correlación de Pearson es muy útil para encontrar correlaciones lineales.
• Si la relación entre las variables NO es lineal, existen otras correlaciones que pueden ser útiles: Spearman y Kendall.
• Si no hay ninguna relación entre dos variables, entonces el coeficiente de correlación será ciertamente 0; sin embargo, si es 0, solo podemos decir que no existe una relación lineal, pero podría existir otra tipo de relación.

Visualización del Conjunto de Datos

La visualización de datos es un aspecto crucial del aprendizaje automático que permite a los analistas comprender y dar sentido a los patrones, relaciones y tendencias de los datos. A través de la visualización de datos, los conocimientos y patrones de los datos se pueden interpretar, lo que los convierte en un componente fundamental del aprendizaje automático.

Veremos la importancia de la visualización de datos en el aprendizaje automático, sus diversos tipos y algunos ejemplos.

Importancia de la visualización de datos en el aprendizaje automático

La visualización de datos ayuda a los analistas de aprendizaje automático a comprender y analizar mejor, conjuntos de datos complejos presentándolos en un formato fácilmente comprensible. La visualización de datos es un paso esencial en la preparación y el análisis de datos, ya que ayuda a identificar valores atípicos, tendencias y patrones en los datos que otras formas de análisis pueden pasar por alto.

Con la creciente disponibilidad de big data, se ha vuelto más importante que nunca utilizar técnicas de visualización de datos para explorar y comprender los datos. Los algoritmos de aprendizaje automático funcionan mejor cuando tienen datos limpios y de alta calidad, y la visualización de datos puede ayudar a identificar y eliminar cualquier inconsistencia o anomalía en los datos.

Tiene tanta importancia seleccionar un buen conjunto de datos casi como seleccionar un buen algoritmo.

Gráficos que trabajan con variables numéricas o cuantitativas

Histograma: El histograma utiliza las famosas tablas de frecuencias. Es un diagrama de barras. La altura de las barras es la frecuencia. Y cada barra se sitúa en su debida clase.

El histograma tiene las barras pegadas.

Gráficos de barras: los gráficos de barras son una forma común de mostrar datos categóricos. En un gráfico de barras, cada categoría está representada por una barra, y la altura de la barra indica la frecuencia o proporción de esa categoría en los datos. Los gráficos de barras son útiles para comparar varias categorías y ver patrones a lo largo del tiempo.

Boxplot o diagramas de caja: los diagramas de caja son una representación gráfica de la distribución de un conjunto de datos. En un diagrama de caja, la mediana se muestra mediante una línea dentro del cuadro, mientras que el cuadro central representa el rango de los datos. Los bigotes se extienden desde el cuadro hasta los valores más altos y más bajos de los datos, excluyendo los valores atípicos. Los diagramas de caja pueden ayudarnos a identificar la dispersión y la asimetría de los datos.

Diagrama de líneas o gráfico de líneas, cada punto de datos está representado por un punto en el gráfico y estos puntos están conectados por una línea. Podemos encontrar patrones y tendencias en los datos a lo largo del tiempo utilizando gráficos de líneas. Los datos de series temporales se muestran con frecuencia mediante gráficos de líneas.

Scatter: El scatter (nube de puntos) es un gráfico de dos variables. El concepto es el mismo que el anterior, pero en lugar de unir los puntos con una línea, se dejan los puntos o crucecitas. Este gráfico es muy útil para intuir cómo se relaciona una variable numérica con otra rápidamente. En la regresión lineal es muy usado para intuir correlaciones o relaciones lineales.

Matrix Plot: Este gráfico se utiliza para graficar 3 o más variables entre sí. Relaciona una variable con las otras en 2D. El diagrama de puntos son scatters 2D de parejas de variables. Además el matrix plot tiene la peculiaridad de poner el histograma para ver la distribución de la variable numérica.

Es muy útil utilizar este tipo de gráficos cuando tienes varias variables numéricas.

Puedes intuir muy rápidamente la relación entre variables.

Mapa de correlaciones: La correlación nos indica la dependencia lineal entre 2 variables.

El mapa de correlaciones nos indica en colorines las variables que están más correlacionadas que las otras.

Cuando el color es más cercano a 1, la correlación es más evidente.

Mapas de calor: los mapas de calor son un tipo de representación gráfica que muestra datos en formato matricial. El valor del punto de datos que representa cada celda de la matriz determina su tono. Los mapas de calor se utilizan a menudo para visualizar la correlación entre variables o para identificar patrones en datos de series temporales.

Histograma + densidad de probabilidad: El histograma nos muestra cómo está distribuida la variable numérica. A esta gráfica le podemos sumar la estimación de la densidad de probabilidad

Los gráficos que trabajan con variables categóricas o nominales

Diagrama de barras: Uno de los gráficos más interesantes es pintar los grupos basados en categorías en forma de barras. En el ejemplo se esta contando cantidad de elementos según su color. Los colores o categorías son seis en este caso.

Diagrama de sectores: Es igual que el ejemplo anterior pero se pinta en forma de pastel. Normalmente se expresa en forma de frecuencia relativa, en proporción, en porcentaje.

Conclusión: el conjunto de datos es una parte esencial en la resolución de un problema

mediante el uso de Machine Learning, es decir, tan importante es que tengamos un conjunto de datos de experiencia pasada suficientemente grande y de calidad como tener o cómo seleccionar un algoritmo que sea adecuado para resolver ese problema.

OverFitting y UnderFitting

Qué es overfitting y underfitting

Las principales causas al obtener malos resultados en Machine Learning son el overfitting (sobreajuste) o el underfitting (subajuste) de los datos. Éstos 2 conceptos hacen referencias a los errores que nuestro modelo puede tener al generalizar el conocimiento que pretendemos que adquieran.

Veamos 2 ejemplos: Supongamos que vemos un perro Labrador por primera vez en la vida y nos dicen “eso es un perro”. Luego nos enseñan un Caniche y nos preguntan: ¿eso es un perro? Diremos “No”, pues no se parece en nada a lo que aprendimos anteriormente.

Ahora imaginemos que nos muestran un libro con fotos de 10 razas de perros distintas. Cuando veamos una raza de perro desconocida, seguramente seremos capaces de

reconocerlo como un cuadrúpedo pero no podremos distinguirlo de un gato o un perro, aunque sea peludo y tenga 4 patas.

Cuando entrenamos nuestros modelos computacionales con un conjunto de datos de entrada estamos haciendo que el algoritmo sea capaz de generalizar un concepto para que al consultarle por un nuevo conjunto de datos desconocido, éste sea capaz de sintetizarlo, comprenderlo y devolvernos un resultado fiable dada su capacidad de generalización.

El problema de la Máquina al Generalizar

• Si nuestros datos de entrenamiento son muy pocos nuestra máquina no será capaz de generalizar el conocimiento y estará incurriendo en underfitting. Este es el caso en el que le enseñamos sólo una raza de perros y pretendemos que pueda reconocer a otras 10 razas de perros distintas. El algoritmo no será capaz de darnos un buen resultado por falta de datos para hacer sólido el conocimiento. También es ejemplo de “subajuste” cuando la máquina reconoce todo lo que “ve” como un perro, tanto una foto de un gato o un coche.

• Por el contrario, si entrenamos a nuestra máquina con 10 razas de perros sólo de color

marrón de manera rigurosa y luego enseñamos una foto de un perro blanco, nuestro modelo no podrá reconocerlo cómo perro por no cumplir exactamente con las características que aprendió (el color forzosamente debía ser marrón). Aquí se trata de un problema de overfitting.

Tanto el problema del ajuste “por debajo” como “por encima” de los datos son malos porque no permiten que nuestra máquina generalice el conocimiento y no nos darán buenas predicciones (o clasificación, o agrupación, etc.)

El equilibrio del Aprendizaje

Deberemos encontrar un punto medio en el aprendizaje de nuestro modelo en el que no estemos incurriendo en underfitting y tampoco en overfitting.

Cuando entrenamos nuestro modelo solemos parametrizar y limitar el algoritmo, por ejemplo la cantidad de iteraciones que tendrá o un valor de “tasa de aprendizaje” (learning-rate) por iteración y muchos otros.

Para lograr que nuestro modelo dé buenos resultados iremos revisando y contrastando nuestro entrenamiento con el conjunto de Test y su tasa de errores,

Para intentar que estos problemas nos afecten lo menos posible, podemos llevar a cabo diversas acciones.

¿Cómo prevenir el Underfitting?

• Una solución al underfitting sería utilizar un modelo, una función hipótesis, un poco más de flexibilidad, es decir, utilizar una función hipótesis que tuviese más características polinómica, que fuese un algoritmo más complejo que la regresión lineal, por ejemplo.

¿Cómo prevenir el Overfitting?

• Aumentar el conjunto de Datos: Obviamente cuanto mayor sea nuestro conjunto de datos, más representativo será, menos overfitting tendremos cuando nos llegue un nuevo dato más probabilidades tendremos de que haya un dato en nuestro conjunto de datos de entrenamiento que se parezca mucho a ese nuevo dato, por lo que, aunque estemos produciendo overfitting es muy probable que se haga una buena predicción.

Clases variadas y equilibradas en cantidad: En caso de aprendizaje supervisado y suponiendo que tenemos que clasificar diversas clases o categorías, es importante que los datos de entrenamiento estén balanceados. Supongamos que tenemos que diferenciar entre manzanas, peras y bananas, debemos tener muchas fotos de las 3 frutas y en cantidades similares. Si tenemos muy pocas fotos de peras, esto afectará en el aprendizaje de nuestro algoritmo para identificar esa fruta.

• Conjunto de datos solo para Test: Siempre subdividir nuestro conjunto de datos y mantener una porción del mismo “oculto” a nuestra máquina entrenada. Esto nos permitirá obtener una valoración de aciertos/fallos real del modelo y también nos permitirá detectar fácilmente efectos del overfitting /underfitting.

• Parameter Tunning o Ajuste de Parámetros: deberemos experimentar sobre todo dando más/menos “tiempo/iteraciones” al entrenamiento y su aprendizaje hasta encontrar el equilibrio.

Reducción del número de características:

Vimos que cuanto más características polinómicas teníamos, más flexible era nuestro modelo y pero también más posibilidades de tener Overfitting tenemos. Con lo cual una de nuestras posibilidades sería poder reducir el número de características.

Se puede observar que cuanto paso de n a 2 características, mi modelo es más rígido pero eliminé el Overfitting. Nosotros podemos reducir el número de características de diferentes formas.

• Selección manual de las características que se deben mantener:

Esta no es una forma muy recomendada porque en muchas ocasiones tendremos un número muy elevado de características de entrada y es muy complejo saber exactamente cuáles eliminar.

• Utilizar un algoritmo para la selección de características:

Podemos utilizar un algoritmo que me indique cuáles características de mi conjunto de datos son importantes y entonces eliminar el resto ya que no van a ser relevantes para construir ese modelo o te van a producir Overfitting.

Uno de los algoritmos más conocidos que podemos utilizar para realizar esta tarea es el Random Forest.

• Utilizar un algoritmo para la extracción de características:

Otra opción es utilizar un conjunto de algoritmos que sirven para la extracción de características.

• Utilizar un algoritmo para la extracción de características (cont.):

Estos tipos de algoritmo se denominan algoritmos de reducción de dimensionalidad y básicamente lo que van a hacer es transformar mi conjunto de datos. Por ejemplo, supongamos que tenemos en nuestro conjunto de datos, 10 características de entrada, luego de utilizar unos de éstos algoritmos, podríamos tener un nuevo conjunto de datos con menos características de entrada. Éstos algoritmos van a transformar el conjunto de datos pues va a modificar también los valores de las características siempre manteniendo la distribución original de los datos.

Algunos de estos algoritmos son, por ejemplo, el algoritmo de Análisis de Componentes

Principales PCA (Principal Component Analysis) o el algoritmo de Descomposición en

Valores Singulares SVD (Singular Value Decomposition).

Regularización:

Es la técnica más utilizada para reducir el Overfitting. Esta técnica viene implementada ya por defecto en todos los algoritmos que implementa la librería Scikit Learn.

Consiste en añadir una penalización a determinadas características de mi modelo, de manera que se reduzca la flexibilidad de mi modelo.

Intuitivamente vemos que cuanto mayor sea la cantidad de características, más flexible va a ser modelo, mejor se va a adaptar a los ejemplos de nuestro conjunto de datos de entrenamiento pero no funcionará correctamente para nuevos datos. Lo que va hacer la regularización es reducir esos valores en aquellos parámetros que están produciendo Overfitting, reduciendo la flexibilidad del modelo, ya no se ajustará tan bien a nuestro conjunto de datos de entrenamiento pero funcionará mejor para nuevos datos.

Vamos a ver más en detalle en qué consiste la regularización.

¿Cómo funciona la regularización?

Cuando vimos la Regresión Lineal, usamos el error cuadrático medio como función de error o coste J:

$J = MSE$

Cuando usamos regularización, le añadimos a la función de error, un término que penaliza la complejidad del modelo. En el caso del MSE, tenemos:

$J = MSE + \alpha.C$

C es la función de penalización, no da una idea de la complejidad del modelo.

El hiperparámetro α (los hiperparámetros son parámetros de un estimador que no se van aprendiendo según se entrena el modelo, si no que son predefinidos antes de empezar el entrenamiento) es el parámetro de regularización.

Cuando usamos regularización minimizamos la complejidad del modelo a la vez que minimizamos la función de coste o error. Esto resulta en modelos más simples que tienden a generalizar mejor ya que los modelos que son excesivamente complejos tienden a tener overfitting.

Regularización Lasso (L1)

En la regularización Lasso, también llamada L1, la complejidad C se mide como la media del valor absoluto de los coeficientes del modelo. Matemáticamente quedaría:

$C = (1/N) \Sigma|wj|, j=1$

Para el caso del error cuadrático medio, este es el desarrollo completo para Lasso (L1):

$J = (1/M) \Sigma(real\_i - estimado\_i)^2 + \alpha(1/N) \Sigma|wi|$

$i=1 \qquad j=1$

¿Cuándo es efectiva Lasso (L1)?

Lasso es útil cuando sospechamos que varios de los atributos de entrada (features) son irrelevantes. Usando Lasso, favorecemos que algunos de los coeficientes sean 0. Esto puede ser útil para descubrir cuáles de los atributos de entrada son relevantes y, en general, para obtener un modelo que generalice mejor. Lasso funciona mejor cuando los atributos no están muy correlacionados entre ellos.

Regularización Ridge (L2)

En la regularización Ridge, también llamada L2, la complejidad C se mide como la media del cuadrado de los coeficientes del modelo. Matemáticamente quedaría:

$C = (1/2N) \Sigma wj^2$

$j=1$

Para el caso del error cuadrático medio, este es el desarrollo completo para Lasso (L1):

$J = (1/M) \Sigma(reali - estimadoi)^2 + \alpha(1/2N) \Sigma wj^2$

$i=1 \qquad j=1$

¿Cuándo es efectiva Ridge (L2)?

Ridge nos va a servir de ayuda cuando sospechamos que varios de los atributos de entrada (features) estén correlados entre ellos. Ridge hace que los coeficientes acaben siendo más pequeños. Esta disminución de los coeficientes minimiza el efecto de la correlación entre los

atributos de entrada y hace que el modelo generalice mejor. Ridge funciona mejor cuando la mayoría de los atributos son relevantes.

Regularización ElasticNet (L1 y L2)

ElasticNet combina las regularizaciones L1 y L2. Con el parámetro r podemos indicar qué importancia relativa tienen Lasso y Ridge respectivamente. Matemáticamente:

C = r · Lasso + (1 - r) · Ridge

¿Cuándo es efectiva ElasticNet?

Usaremos ElasticNet cuando tengamos un gran número de atributos. Algunos de ellos serán irrelevantes y otros estarán correlados entre ellos.

Comparación Ridge y Lasso

La principal diferencia práctica entre lasso y ridge es que el primero consigue que algunos coeficientes sean exactamente cero, por lo que realiza selección de predictores, mientras que el segundo no llega a excluir ninguno. Esto supone una ventaja notable de lasso en escenarios donde no todos los predictores son importantes para el modelo y se desea que los menos influyentes queden excluidos.

Por otro lado, cuando existen predictores altamente correlacionados (linealmente), ridge reduce la influencia de todos ellos a la vez y de forma proporcional, mientras que lasso tiende a seleccionar uno de ellos, dándole todo el peso y excluyendo al resto. En presencia de correlaciones, esta selección varía mucho con pequeñas perturbaciones (cambios en los datos de entrenamiento), por lo que, las soluciones de lasso, son muy inestables si los predictores están altamente correlacionados.

Para conseguir un equilibrio óptimo entre estas dos propiedades, se puede emplear lo que se conoce como penalización elastic net, que combina ambas estrategias.

Evaluación de la Función Hipótesis

¿Cómo podemos saber si un número elevado de características provoca Overfitting?

Se requiere una forma de evaluar la función hipótesis generada para comprobar si generaliza correctamente:

Una o dos características: Se pueden representar gráficamente la función hipótesis.

Para más de dos características:

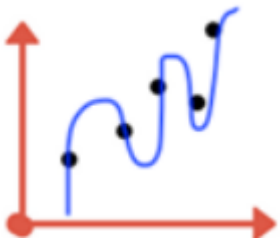
Lo que haremos será dividir nuestro conjunto de datos en diferentes subconjuntos y utilizar cada uno de ellos para una tarea específica, es decir, para entrenar nuestro algoritmo utilizaremos un subconjunto y para evaluar que nuestro algoritmo se comporta correctamente, utilizaremos otros subconjuntos.

Veamos esto con más detalle:

Imagina que este nuestro conjunto de datos de entrenamiento. Observemos que solo tenemos una sola característica de entrada “x” y una característica de salida “y”

Podríamos pasarle en conjunto de datos a una algoritmo de machine learning, por ejemplo, uno de regresión polinómica y construir una función hipótesis.

Que pasa si la función obtenida de la regresión polinómica casi que no deja margen de error la curva cae casi en puntos del entrenamiento. Si calculamos el error nos podría dar un valor tan bajo que a primera instancia obtendremos un modelo perfecto.



Ahora si usamos esta misma función para predecir con los datos diferentes, nos podremos encontrar con un error más alto esto es la que llamamos Overfitting

Es decir, que si yo entreno todo mi algoritmo, es decir, construyo una función hipótesis con todo mi conjunto de datos, luego no tengo forma de saber si está produciendo Overfitting o no, porque si yo me quedo únicamente con esta característica de entrada y la evalúo sobre mi función hipótesis, obviamente, como ésta pasando por todas mis características de entrada, me va a dar una predicción perfecta y voy a pensar que ese modelo se comporta de manera perfecta.

Lo voy a poner en producción y cuando me llegue un nuevo ejemplo, que no ha visto nunca va a dar una mala predicción. Entonces ¿qué hacemos?

Train y Test

Para encontrar este tipo de problemas, lo que vamos a hacer es dividir nuestro conjunto de datos en dos subconjuntos, un primer subconjunto llamado conjunto de entrenamiento “Train Set” y otro subconjunto llamado conjunto de pruebas “test set”.

Por lo general el “Train Set” se lo dimensiona alrededor del 60% o 70% y el “Test Set” será un 40% o 30% de nuestros conjunto de datos.

Y, ahora entonces ¿que vamos a hacer?

Ahora, vamos a utilizar el subconjunto “Train Set” para entrenar mi algoritmo, voy a generar mi modelo, mi función hipótesis. Podemos observar que se produjo Overfitting.

Para probar el modelo no voy a utilizar este mismo subconjunto, ya que como vimos, la predicción sería perfecta, sino que voy a utilizar el segundo subconjunto, el “Test Set”. Pero de este subconjunto, solo tomaré las características “x” y veré la predicción “y” que responde mi modelo. Una vez calculadas todas las predicciones del “Test Set”, lo que hago es comparar esas predicciones con la etiqueta que tienen y que sé que es el valor real de estos ejemplos con los que yo he predecido.

Conclusión

• Si el modelo entrenado con el conjunto de entrenamiento tiene un 90% de aciertos y con el conjunto de test tiene un porcentaje muy bajo, esto señala claramente un problema de overfitting.

• Si en el conjunto de Test sólo se acierta un tipo de clase (por ejemplo “peras”) o el único resultado que se obtiene es siempre el mismo valor será que se produjo un problema de underfitting.

Evaluación de Resultados

¿Qué es una Métrica?

En clasificación son números que miden el rendimiento de un modelo de aprendizaje automático cuando se trata de asignar observaciones a ciertas clases. Generalmente, el rendimiento se presenta en un rango de 0 a 1, donde la puntuación 1 corresponde a un modelo perfecto, clasificado correctamente.

Tenemos distintas Métricas dependiendo si tenemos un problema de:

• Regresión

• Clasificación.

Métricas de Regresión

En análisis de regresión las principales métricas son las siguientes:

• Error cuadrático medio o Mean Squared Error (MSE)
• Error absoluto medio o Mean Absolute Error (MAE)
• R al cuadrado (R2)

Otra métricas son:

• Error cuadrático medio (RMSE)
• R cuadrado ajustado (R2)
• Error de porcentaje cuadrático medio (MSPE)
• Error porcentual absoluto medio (MAPE)
• Error logarítmico cuadrático medio (RMSLE)

Error cuadrático medio o Mean Squared Error (MSE)

Se define como el valor medio de los cuadrados de la diferencia entre los valores predichos y los valores reales:

$MSE = (1/n) * \Sigma(y_n - \bar{y})^2$

Donde:

- y_n son los valores reales

- ȳ son los valores predichos

- n es el número de observaciones

python

from sklearn.metrics import mean_squared_error

mean_squared_error([2, 5, 9], [3, 5, 11])

1.6666666666666667

Error absoluto medio o Mean Absolute Error (MAE)

Se define como el valor medio de la suma de los valores absolutos de la diferencia entre los valores predichos y los valores reales:

mae = (1/n) Σ |yi - ŷi|

Donde:

- mae representa el Error Absoluto Medio
- n es el número de observaciones
- yi representa los valores reales
- ŷi representa los valores predichos
- Σ indica la sumatoria
- |yi - ŷi| representa el valor absoluto de la diferencia entre el valor real y el valor predicho

El MAE es una métrica común para evaluar el rendimiento de modelos predictivos, midiendo la magnitud promedio de los errores sin considerar su dirección.

python

from sklearn.metrics import mean_absolute_error

mean_absolute_error([2, 5, 9], [3, 5, 11])

1.0

R al cuadrado (R^2)

R^2 determina la capacidad de un modelo para predecir futuros resultados. R^2 siempre estará entre -∞ y 1. El mejor resultado posible es 1 y ocurre cuando la predicción coincide con los valores de la variable objetivo. R2 puede tomar valores negativos pues la predicción puede ser arbitrariamente mala. Cuando la predicción coincide con la esperanza de los valores de la variable objetivo, el resultado de R2 es 0. Se calcula con esta fórmula:

$R^2 = 1 - \Sigma(yi - pi)^2 / \Sigma(yi - \bar{y})^2$

donde: yi son los valores que toma la variable objetivo pi son los valores de la predicción $\bar{y}$ es el valor medio de los valores que toma la variable objetivo Scikit-Learn implementa esta métrica en sklearn.metrics.r2_score:

from sklearn.metrics import r2_score r2_score([2, 5, 9], [3, 5, 11]) 0.7972972972972973"

Esta métrica indica qué proporción de la varianza de la variable objetivo es explicada por el modelo, siendo 1 el ajuste perfecto y valores negativos indicando un modelo peor que usar simplemente la media.

Métricas de Clasificación

Básicamente existen diferentes formas de evaluar los resultados de un algoritmo de clasificación:

• Métricas numéricas: Matriz de Confusión. Accuracy (Exactitud). Precision (Precisión). Recall (Exhaustividad), F1 Score

• Representaciones gráficas: Curva ROC, Curva PR

Matriz de Confusión

También conocida como matriz de error, es una tabla resumida que se utiliza para evaluar el rendimiento de un modelo de clasificación. El número de predicciones correctas e incorrectas se resumen con los valores de conteo y se desglosan por cada clase.

Vemos una matriz de confusión de 2x2:

Supongamos dos clases: positiva y negativa.

Verdadero Positivo (VP): Resultado en el que el modelo predice correctamente la clase positiva.

Verdadero Negativo (VN): Resultado donde el modelo predice correctamente la clase negativa.

Falso Negativo (FN): es la cantidad de positivos que fueron clasificados incorrectamente como negativos.

Falso Positivo (FP): es la cantidad de negativos que fueron clasificados incorrectamente como positivos.

	Predicción	
	Positivos	**Negativos**
Observación Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
Observación Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Matriz de Confusión Métricas

Accuracy (Exactitud)

La exactitud mide el porcentaje de casos que el modelo ha acertado.

Se calcula de la siguiente manera:

Accuracy = [(VP + VN) / (VP+FP+FN+VN)] ×100%

$$\text{Accuracy} = \frac{990 + 0}{990 + 0 + 10 + 0}$$

La Accuracy del modelo es básicamente el número total de predicciones correctas dividido por el número total de predicciones. En este caso da 99%, pero no pudo identificar ningún perro. La visión que me da esta métrica no es completa.

```
from sklearn.metrics import accuracy_score
accuracy = accuracy_score(y_true, y_pred)
```

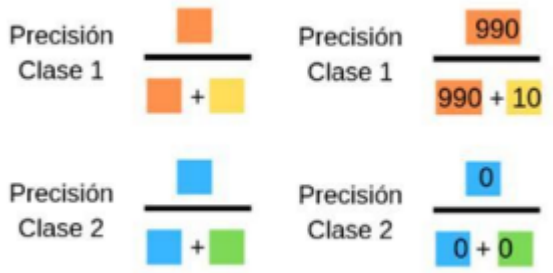
Precision (Precisión)

Se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión.

En forma práctica es el porcentaje de casos positivos detectados.

La Precisión de una clase define cuan confiable es un modelo en responder si un punto pertenece a esa clase. Para la clase gato será del 99% sin embargo para la de perro será 0%.

Precision = VP / (VP + FP)



from sklearn.metrics import precision_score

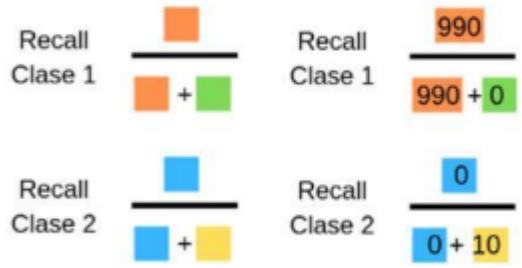
precision = precision_score(y_true, y_pred)

Recall (Exhaustividad)

La métrica de exhaustividad nos va a informar sobre la cantidad que el modelo de machine learning es capaz de identificar. El Recall de una clase expresa cuán bien puede el modelo detectar a esa clase. Para gatos será de 1 y para perros 0.

Para calcular la exhaustividad (recall) usaremos la siguiente fórmula:

Recall = VP / (VP + FN)



from sklearn.metrics import recall_score

recall = recall_score(y_true, y_pred)

F1 Score

Esta es otra métrica que nos resume la precisión y el recall (exhaustividad) en una sola métrica. La fórmula de F1 Score es la media armónica de la precisión y el recall. Un modelo perfecto tiene una puntuación F de 1 y muy malo de cero 0.

En nuestro caso nos da cero para perros.

La imagen muestra la fórmula matemática para calcular el F1 Score:

F1 Score = 2 / (1/Precisión + 1/Recall) = (2 * Precisión * Recall) / (Precisión + Recall)

Esta fórmula representa el F1 Score como la media armónica de la precisión y el recall, lo que proporciona un balance entre ambas métricas. Es especialmente útil cuando las clases están desbalanceadas, ya que considera tanto los falsos positivos como los falsos negativos.

from sklearn.metrics import f1_score

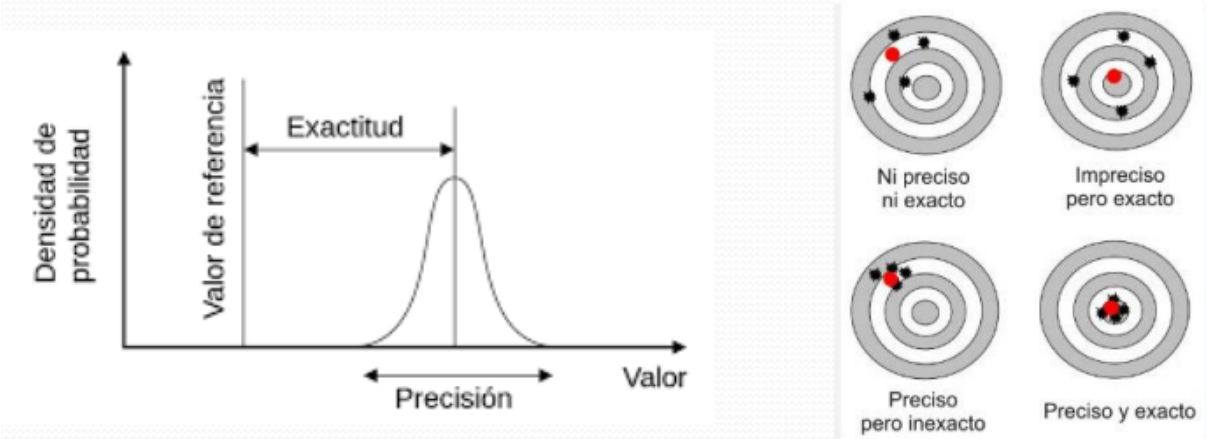
f1 = f1_score(y_true, y_pred)

Parte 2

¿Cuál es la diferencia entre Accuracy y Precision?

Exactitud se refiere a cuán cerca del valor real se encuentra el valor medido. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación. Cuanto menor es el sesgo más exacta es una estimación.

Precisión se refiere a la dispersión del conjunto de datos estimados. Cuanto menor es la dispersión mayor la precisión. En términos estadísticos, la precisión esta relacionada con la desviación estándar de las mediciones.



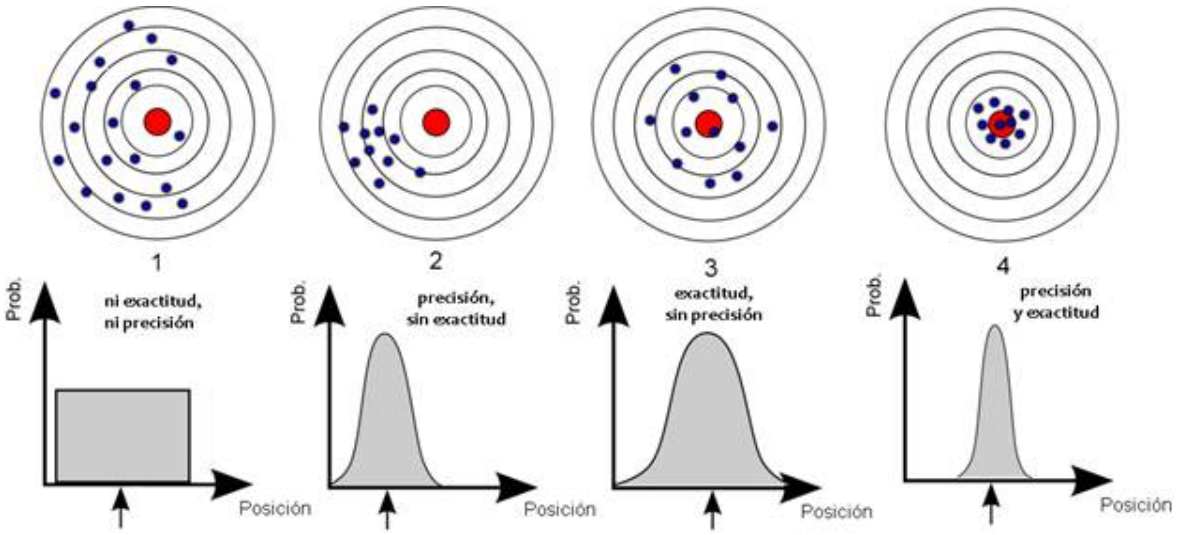
Relación entre Accuracy (Exactitud), Precision (Precisión), Sesgo y Dispersión

Caso1: Baja precisión y exactitud: El modelo no logra clasificar la clase correctamente.

Caso2: Alta precisión y baja exactitud: El modelo no detecta la clase muy bien, pero cuando lo hace es altamente confiable.

Caso3: Baja precisión y alta exactitud: El modelo detecta bien la clase, pero también incluye muestras de la otra clase.

Caso4: Alta precisión y exactitud: el modelo maneja perfectamente esa clase.



Clases desequilibradas - Paradoja de la Exactitud

Supongamos que tenemos un sistema para predecir si las transacciones con tarjeta de crédito son fraudulentas o no.

El resultado es una variable categórica con dos clases : Sí y No.

A medida que recopila datos del mundo real, observa que sólo un pequeño porcentaje (digamos el 3%) de las transacciones son fraudulentas. Por lo tanto, un conjunto de datos con 10000 observaciones tendrá los siguientes recuentos de clases de salida:

Producción	contar
Si(fraudulento)	300
No(genuino)	9700

Cuando la variable de salida tiene una disparidad tan amplia en la frecuencia de clases, decimos que la salida tiene clases desequilibradas y que el conjunto de datos está desequilibrado.

Si utilizamos este conjunto de datos para entrenar un modelo de clasificación, es posible que el modelo no tenga suficientes observaciones con el resultado Sí para aprender a identificar transacciones fraudulentas.

Supongamos que entrenamos un modelo de clasificación para detectar transacciones fraudulentas y debemos evaluar su desempeño.

Tomamos un subconjunto de datos (train Set) de 1000 observaciones en donde también contiene un 3% de casos fraudulentos:

Producción	contar
Si(fraudulento)	30
No(genuino)	970

Luego de entrenar el modelo podemos resumir los resultados en la siguiente matriz de confusión:

El modelo predijo correctamente todas las transacciones genuinas. Así tenemos 970 Verdaderos Negativos y 0 Falsos Positivos. Sin embargo, el modelo solo predijo correctamente una transacción fraudulenta. Entonces generó 29 Falsos Negativos y 1 Verdadero Positivo.

La Exactitud del modelo es:

Exactitud = 971/1000 = 0.971

El modelo tiene una exactitud del 97,1%, aunque no pudo detectar 29 de 30 transacciones fraudulentas.

Conclusión:

Para conjuntos de datos con clases desequilibradas, un modelo podría lograr una alta exactitud al predecir la clase mayoritaria la mayor parte del tiempo .

Por lo tanto, una alta exactitud podría ser engañosa y no garantiza un modelo con buen rendimiento. Este fenómeno aparentemente contradictorio se conoce como paradoja de la

exactitud.

Calculemos la Precisión, Recuperación y Puntuación F1 del modelo.

La Precisión del modelo:

La precisión responde a la siguiente pregunta: ¿Qué porcentaje de todas las predicciones positivas hechas por el modelo fueron correctas?

Precision = 1 / 1+0 = 1

La precisión es una métrica útil cuando se desea minimizar los falsos positivos .

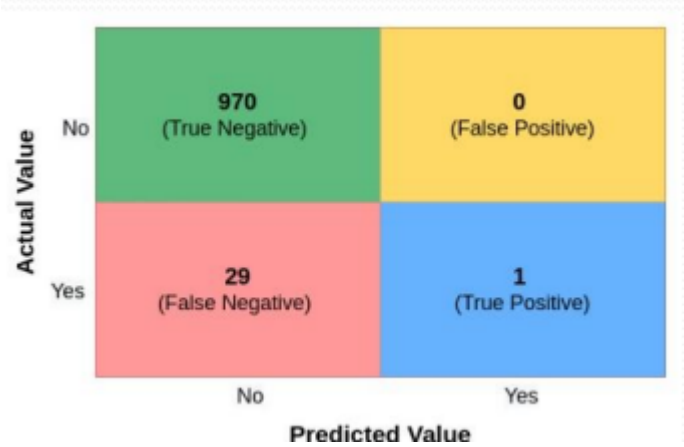
Por ejemplo, un banco que ofrece préstamos. No desea aprobar un préstamo (Positivo) para alguien que no podrá pagarlo. En realidad, su préstamo no debería ser aprobado (Negativo).

En tales casos, debes buscar un modelo con alta precisión .

El Recall del modelo es:

Definimos Recuperar de la siguiente manera: ¿Qué porcentaje de todos los Positivos reales fueron predichos con precisión por el modelo?

recall = 1 / 30 = 0.033



El Recall es una métrica útil cuando desea minimizar los falsos negativos .

Por ejemplo, imagine que está probando si un paciente está infectado con un virus peligroso. No desea un modelo que prediga que un paciente no está infectado (Negativo) cuando el paciente tiene el virus (Positivo).

En tales casos, debes buscar un modelo con un alto Recall .

El F1 Score del modelo es:

F1 Score = (2 * 1 * 0.333) / (1+0.333) = 0.065

Idealmente, queremos construir un modelo con alta Precisión y un alto Recall. Pero, por lo general, hay una compensación: intentar aumentar la Precisión reducirá el Recall y viceversa.

La puntuación F1 se define como la media armónica de precisión y recuperación . Si alguno de ellos llega a ser extremadamente bajo, la puntuación F1 también bajará. Por lo tanto, F1 Score puede ayudarlo a encontrar un buen equilibrio entre Precisión y Recall.

Con los resultados obtenidos podemos concluir que el modelo tiene alta exactitud y precisión. Pero los puntajes extremadamente bajos de Recall y F1 Score indican que tenemos un modelo muy malo.

Específicamente, la puntuación de Recall nos dice que nuestro modelo puede detectar solo el 3,3% de las transacciones fraudulentas.

Relación entre Accuracy - Precision - Recall

La Exactitud (que mide el % de casos que el modelo ha acertado sin distinguir ninguna clase) para juzgar un modelo, es válida en un modelo de clasificación binaria sólo si los datos están balanceados (igual población en ambas clases). Esto se debe a que cuando el 90% de los datos están en clase 0; si el modelo clasifica todos los elementos de entrada como clase 0, la precisión es del 90 % (ya que el modelo se equivoca solo el 10 % de las veces). Esto es muy engañoso.

En realidad, queremos medir de la población que es realmente de la clase 1, qué porcentaje se clasificó como clase 1 (de manera similar para la clase 0); y de la población clasificada por el modelo como clase 1, qué % eran en realidad clase 1 (de manera similar para la clase 0).

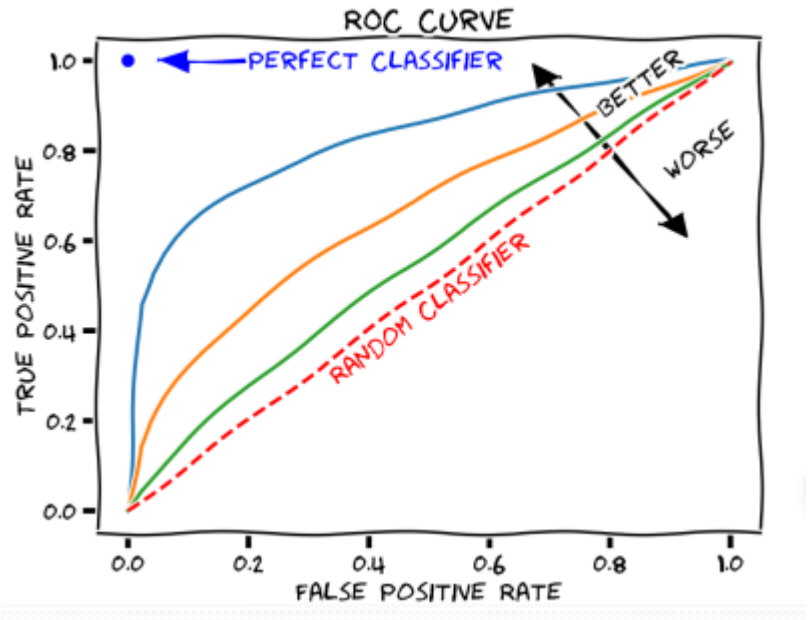
Por lo tanto, las mejores métricas para medir el rendimiento son Precisión y Recall.

La curva ROC (Receiver Operating Characteristic) nos ayuda a mejor la visión de la métrica. Es un gráfico entre la tasa de verdaderos positivos y la tasa de falsos positivos.

Curvas ROC (Receiver Operating Characteristic)

La curva ROC se utiliza para evaluar el rendimiento de los algoritmos de clasificación binaria, es decir, entre dos clases o categorías (1 o 0, Verdadero o Falso, etc.). La curva ROC proporciona una representación gráfica, en lugar de un valor único como la mayoría de las otras métricas.

Se genera calculando la tasa de verdaderos positivos (cuantos gatos reales se predicen como gatos) contra la tasa de falsos positivos (cuantos gatos reales se predicen como perros) de un solo clasificador en una variedad de umbrales. Por ejemplo, en un modelo de regresión logística, si se predice que una observación será positiva con una probabilidad mayor que 0.5, entonces se etiqueta como positiva. Sin embargo, realmente podríamos elegir cualquier umbral entre 0 y 1 (0.1, 0.3, 0.6, 0.99, etc.), y, de este modo, la curva ROC nos ayudaría a visualizar cómo estas elecciones afectan al rendimiento del clasificador.



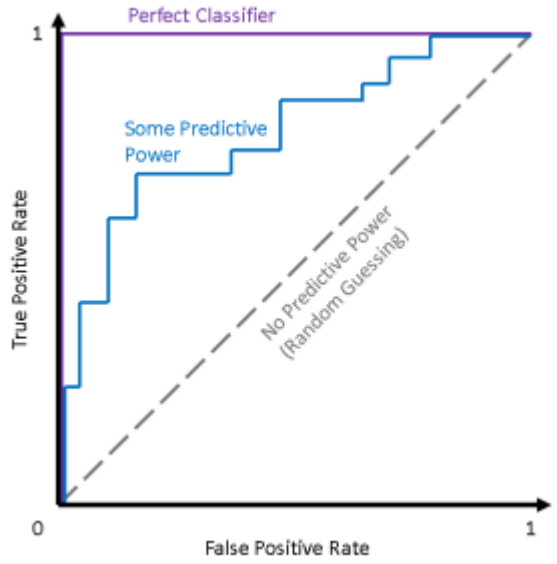
En la siguiente figura se muestra como sería la curva ROC para tres hipotéticos modelos de clasificación distintos:

La línea de puntos gris representa un clasificador aleatorio (equivalente a “adivinar”); su curva ROC consiste en una línea recta diagonal de pendiente 1.

La línea violeta representa un clasificador perfecto, es decir, que no comete ningún error (uno con una tasa de verdaderos positivos del 100% y una tasa de falsos positivos del 0%).

Todos los modelos de ML se ubicarán en algún lugar entre estas dos líneas (línea azul).

Lo que buscamos es un clasificador que mantenga una alta tasa de verdaderos positivos y al mismo tiempo tenga una baja tasa de falsos positivos.

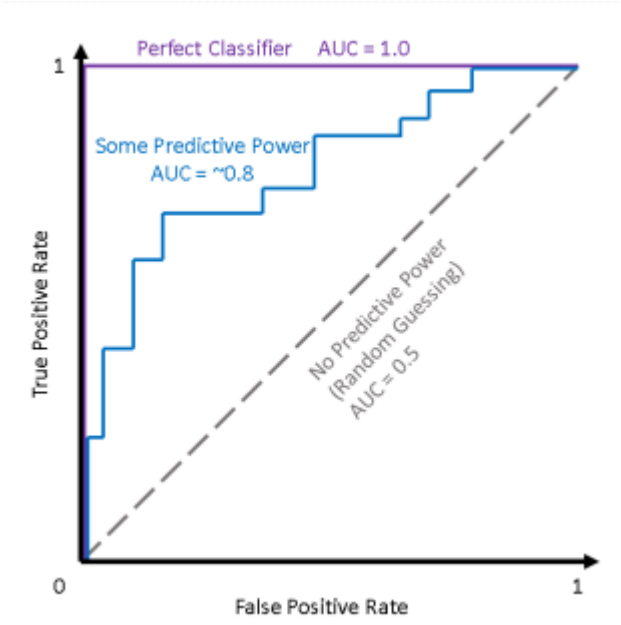


AUC "Area under the ROC Curve“

La curva ROC es muy útil visualmente, pero podemos resumir esta información en una sola métrica: el AUC. Cuanto mayor es la puntuación AUC, mejor es el rendimiento de un clasificador binario.

La siguiente figura muestra que para un clasificador sin poder predictivo (es decir, de predicción aleatoria), el AUC es 0.5, y para un clasificador perfecto, el AUC es 1.0.

La mayoría de los clasificadores estarán entre 0.5 y 1.0, con la rara excepción de que el clasificador funcione peor que la predicción aleatoria (AUC<0.5); en este caso se estarían efectuando predicciones invertidas, es decir, se tendería a predecir la clase negativa cuando está fuera realmente positiva y viceversa.

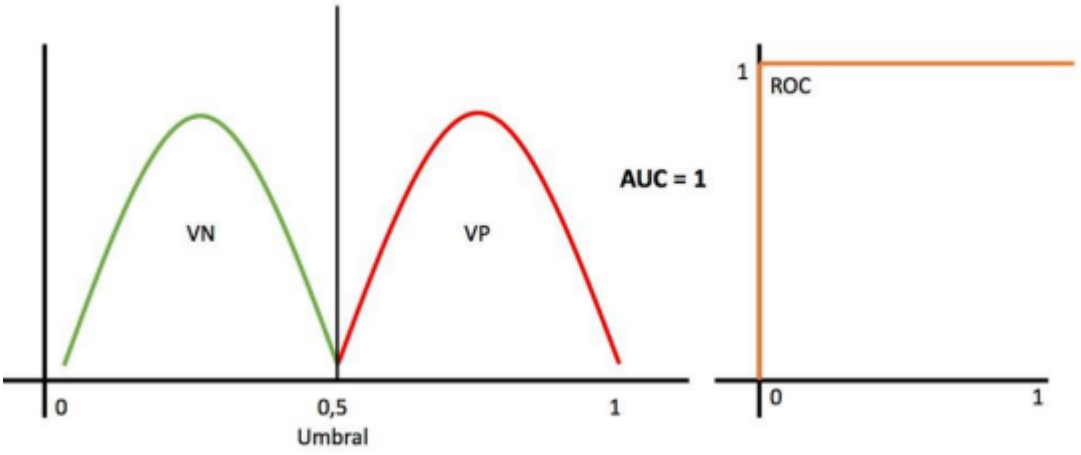


Curva ROC y AUC

Como ya se dijo el AUC es el área bajo la curva ROC. Este puntaje nos da una idea de qué tan bien funciona el modelo. Veamos algunos ejemplos de esto.

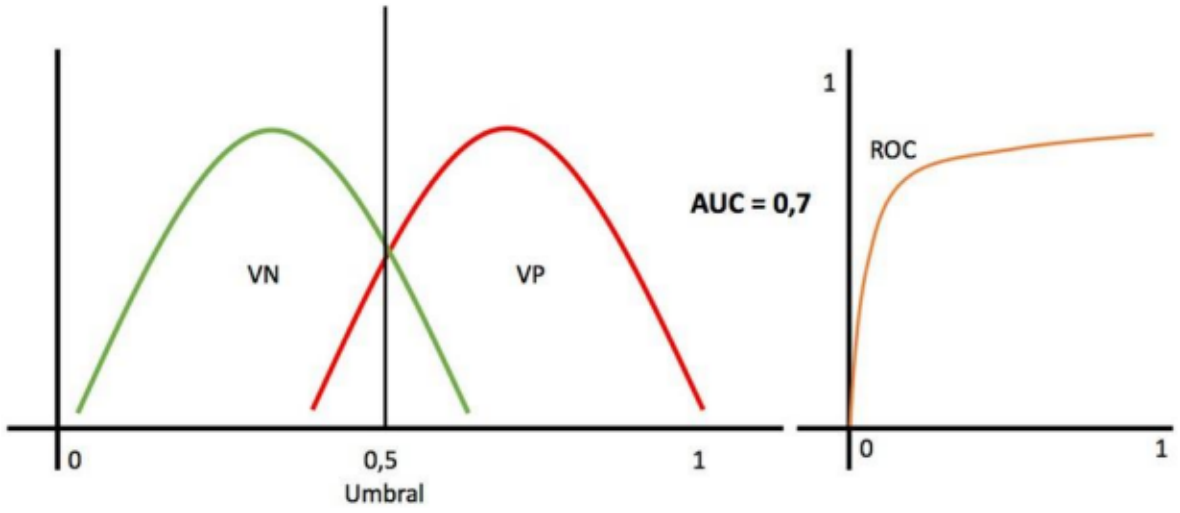
Ejemplo N°1:

Esta es una situación ideal. Cuando dos curvas no se superponen en absoluto, el modelo tiene una medida ideal de separación. Es perfectamente capaz de distinguir entre clase positiva y clase negativa.



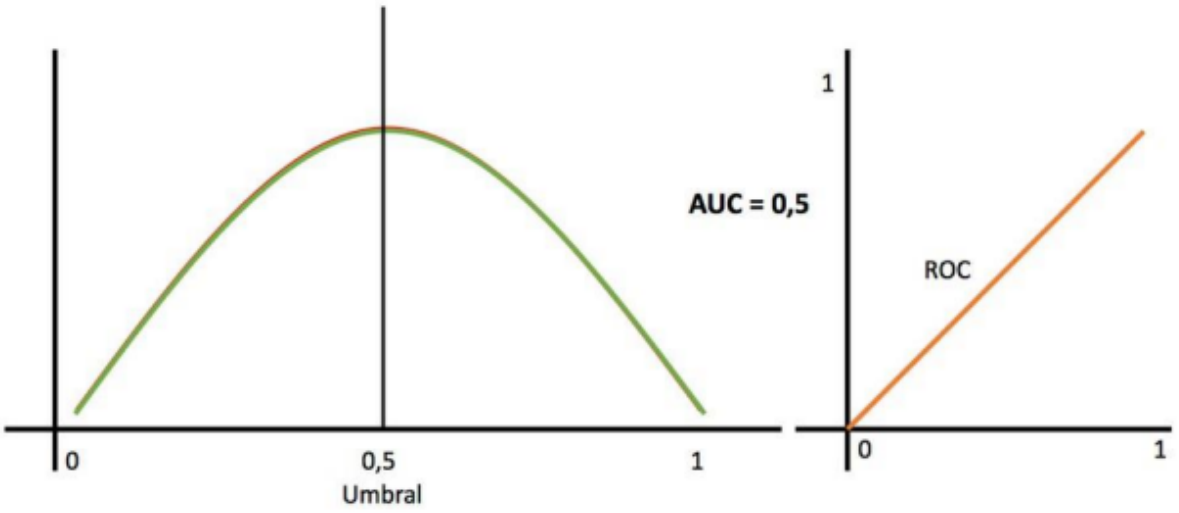
Ejemplo N°2:

Cuando dos distribuciones se superponen, introducimos errores. Dependiendo del umbral, podemos minimizarlos o maximizarlos. Cuando AUC es 0.7, significa que hay 70% de probabilidad de que el modelo pueda distinguir entre clase positiva y clase negativa.



Ejemplo N°3:

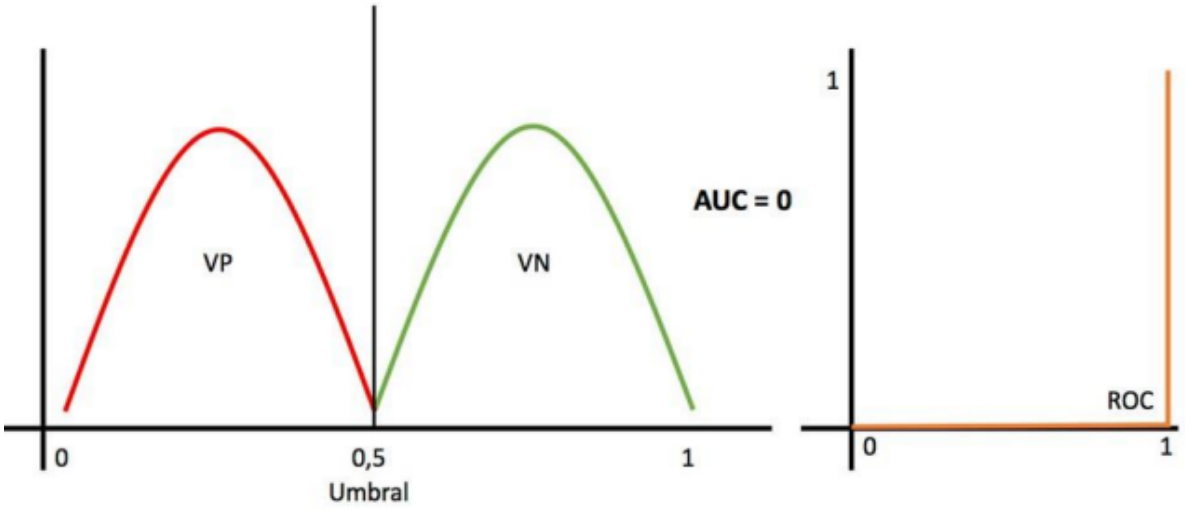
Esta es la peor situación. Cuando el AUC es aproximadamente 0.5, el modelo no tiene capacidad de discriminación para distinguir entre clase positiva y clase negativa.



Ejemplo N°4:

Cuando AUC es aproximadamente 0, el modelo en realidad está correspondiendo las clases.

Significa que el modelo predice la clase negativa como una clase positiva y viceversa.



Curva PR

• La curva ROC proporciona una imagen demasiado optimista del rendimiento, cuando se trata de una clasificación desequilibrada, ya que relaciona la sensibilidad (Tasa de verdaderos positivos) con el porcentaje de falsos positivos y entonces si aumentamos la sensibilidad, el modelo será más optimista, es decir, clasificará más datos como positivos y por tanto, aumentarán los falsos positivos.

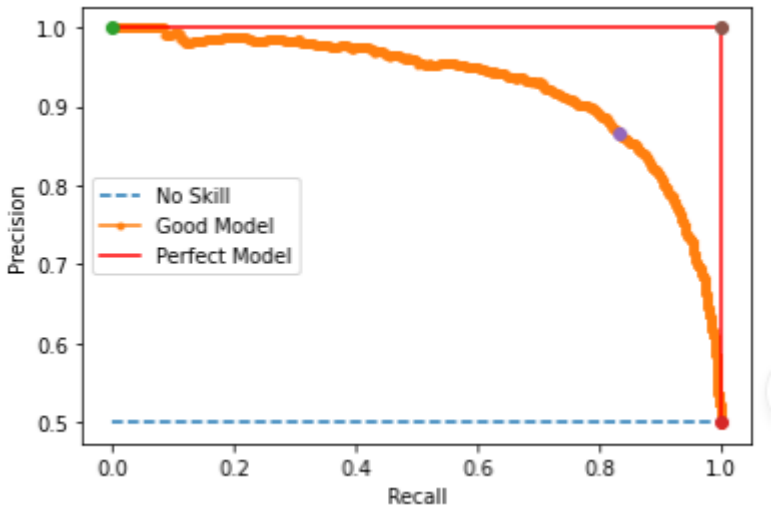
• Pero me pregunto, la curva ROC relaciona la tasa de verdaderos positivos y la tasa de falsos positivos, ¿y que pasa con la clase negativa?.

Recordemos:

La Precisión de una clase define cuán confiable es un modelo en predecir si un punto pertenece a una clase dada. La mejor Precisión que puede tener un modelo es = 1 para cada clase.

El Recall de una clase expresa cuan bien puede el modelo detectar a esa clase real (mejor valor = 1).

La línea roja corresponde a una modelo ideal.



Cómo balancear un Datasets desbalanceado

Recopilar más datos

Verifica si es posible reunir más datos para el problema, un conjunto de datos más grande podría exponer una perspectiva diferente y quizás más equilibrada de las clases.

Utilizar las métricas de evaluación correctas

Vimos que la métrica de Exactitud no es confiable para datasets desbalanceados. Podríamos usar otras métricas de evaluación, tales como:

Precisión y Especificidad: Cuántas instancias seleccionadas son relevantes

Sensibilidad: Cuántas instancias relevantes se seleccionan

Puntaje de F1: Media armónica de precisión y sensibilidad

AUC: Relación entre tasa de verdaderos positivos y tasa de falsos positivos

Refinamiento de algoritmos:

Ajustar los parámetros de los clasificadores para que tengan en cuenta el desbalanceo existente y/o modificar la pesos de las clases para equilibrarlas.

Remuestreo del conjunto de datos

Consiste en trabajar para obtener diferentes conjuntos de datos. La idea principal de las clases de muestreo es aumentar las muestras de la clase minoritaria o disminuir las muestras de la clase mayoritaria.

Puede haber dos tipos principales de muestreo:

1. Puede eliminar instancias de la clase mayoritaria, lo que se denomina Undersampling.

2. Puede agregar copias de instancias de la clase minoritaria, lo que se denomina Oversampling.

Técnicas de Preprocesado – Undersampling:

Eliminar aleatoriamente instancias de la clase de la mayoría de un conjunto de datos, se conoce como Undersampling.

Ventajas:

Al reducir el número de muestras, puede ayudar a mejorar el tiempo de ejecución del modelo y uso de memoria, especialmente cuando el conjunto de datos de entrenamiento es muy grande.

Desventajas:

Puede descartar información útil. La muestra elegida por Undersampling puede ser una muestra sesgada, que no será representativa de la población. Por lo tanto, puede hacer que el clasificador se comporte mal en datos reales que no se ven.

Aplicar un Undersampling cuando tengas muchos datos.

Técnicas de Preprocesado – Oversampling:

Aumenta las instancias correspondientes a la clase minoritaria, sin reducir el número de instancias de la clase mayoritaria se denomina Oversampling.

Supongamos un conjunto de datos con 1000 instancias donde 980 instancias corresponden a la clase mayoritaria y las 20 instancias a la clase minoritaria. Podríamos copiar muestras de la clase minoritaria.

Ventajas:

Este método no conduce a la pérdida de información.

Desventajas:

Se replican eventos de la clase minoritaria artificialmente. Aplicar un sobre-muestreo cuando no tengas muchos datos.

Parte 3

Validación de modelos predictivos

Conjunto de entrenamiento (training set): datos observaciones con las que se entrena el modelo.

Conjunto de validación y conjunto de test (validation set y test set): datos/observaciones del mismo tipo que las que forman el conjunto de entrenamiento pero que no se han empleado en la creación del modelo. Son datos que el modelo no ha “visto”.

Error de entrenamiento (training error): error que comete el modelo al predecir observaciones que pertenecen al conjunto de entrenamiento.

Error de validación y error de test (evaluation error y test error): error que comete el modelo al predecir observaciones del conjunto de validación y del conjunto de test. En ambos casos son observaciones que el modelo no ha “visto”.

Introducción

La finalidad de un modelo es predecir la variable respuesta en observaciones que el modelo no ha “visto” antes.

El error que obtenemos luego de entrenar un modelo es el error de entrenamiento, el error que comete el modelo al predecir las observaciones utilizando los mismo datos que en la fase de entrenamiento, es decir, que ya ha “visto”, no me indican como se comportará el modelo ante nuevas observaciones. Para conseguir una estimación más certera, se tiene que recurrir a estrategias de validación basadas en resampling.

La idea en la que se basan éstos métodos es la siguiente:

El modelo se ajusta empleando un subconjunto de observaciones del conjunto de entrenamiento y se evalúa (calcular una métrica que mida cómo de bueno es el modelo, por ejemplo, accuracy) con las observaciones restantes.

Este proceso se repite múltiples veces y los resultados se agregan y promedian. Gracias a las repeticiones, se compensan las posibles desviaciones que puedan surgir por el reparto aleatorio de las observaciones. La diferencia entre éstos métodos suele ser la forma en la que se generan los subconjuntos de entrenamiento/validación.

Métodos de remuestreo (resampling)

Los métodos de remuestreo se basan en extraer muestras repetidamente a partir de un set de datos de entrenamiento, ajustando el modelo de interés para cada muestra. Se trata de métodos no paramétricos, que no requieren ninguna asunción sobre la distribución de la población.

Dos de los métodos más utilizados de remuestreo son la validación cruzada y el bootstrap:

Validación cruzada: puede aplicarse para estimar el test error asociado a un determinado método de aprendizaje estadístico (tanto regresión como clasificación) para evaluar el rendimiento del modelo, pero también para seleccionar niveles apropiados de flexibilidad, como el grado de polinomio, etc. O seleccionar el modelo usando el test error.

Bootstrap: se lo suele usar para el cálculo estadístico, evaluación de la precisión (error estándar, intervalos de confianza...).

Validación Cruzada (cross-validation)

Estimar el test error (no conocemos el verdadero test error) excluyendo una parte de las observaciones en el proceso de ajuste del modelo, usando luego dichas observaciones para evaluar la capacidad predictiva del modelo.

La diferencia principal entre un método u otro es la forma en la que se generan los grupos de entrenamiento y test.

Método de retención o Validación simple (holdout method)

Este enfoque es el denominado Train-Test Split consiste en descomponer de manera aleatoria una serie de datos. Una parte servirá para el entrenamiento del modelo de Machine Learning, la otra permitirá probarlo para la validación.

Por lo general, se reserva entre un 70 % y 80 % de los datos de la serie para el entrenamiento. El 20-30 % restante se reserva para Test.

Esta técnica es eficaz, si los datos no son escasos, en este caso puede faltar información contenida en los datos que no se utilizan para el entrenamiento y, por tanto, los resultados pueden tener un gran sesgo. Por el contrario, si los datos son suficientes y la distribución es uniforme entre las dos muestras, este enfoque es adecuado.

Se pueden separar los datos de manera manual o usar el método train_test split de scikit-learn.

Tiene dos problemas importantes:

• La estimación del error es altamente variable dependiendo de qué observaciones se incluyan como conjunto de entrenamiento y cuáles como conjunto de validación (problema de varianza).

• Al excluir parte de las observaciones disponibles como datos de entrenamiento (generalmente el 20%), se dispone de menos información con la que entrenar el modelo y, por lo tanto, se reduce su capacidad de predecir correctamente. Esto suele tener como consecuencia una sobrestimación del error comparado al que se obtendría si se emplearan todas las observaciones para el entrenamiento (problema de bias).

K-Fold Cross-Validation

Consiste en dividir los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño, k-1 grupos se emplean para entrenar el modelo y uno de los grupos se emplea como validación. Este proceso se repite **k** veces utilizando un grupo distinto como validación en cada iteración.

El proceso genera **k** estimaciones del error cuyo promedio se emplea como estimación final.

Es importante notar, que cada dato aparece una sola vez en los datos de prueba y k-1 en los datos de Entrenamiento.

Ventajas del método K-Fold Cross-Validation

• El hiperparámetro más importante es k que se refiere al número de grupos en que se dividirá una muestra de datos dada.

• Es un método popular porque es fácil de entender y porque generalmente resulta en una estimación menos sesgada o menos optimista de la habilidad del modelo que otros métodos, como la validación simple (train test split).

• Un valor común de k que puede dar buenos resultados en cuanto al equilibrio bias-varianza para estimar el test error rate es K = 5 o K = 10.

Puedo comparar muchos modelos (el mismo algoritmo, por ejemplo, regresión lineal con distintos valores de los parámetros m y b).

Hago 4 mezclas de los datos, elijo en valor de K y en cada iteración (1°fold, 2°fold, etc) calculo la performance de cada modelo (un algoritmo de regresión lineal con 4 modelos, 4 valores de m y b).

Cada clasificador es un modelo distinto.

Leave-group-out Cross Validation (LGOCV) – Validación cruzada aleatoria

También conocido como Monte Carlo CV, separa de manera aleatoria los subgrupos de datos entrenamiento/test (cuyo porcentaje se establece de antemano) múltiples veces.

Leave One Out Cross-Validation (LOOCV)

Implica separar los datos de forma que para cada iteración tengamos una sola muestra para los datos de prueba y todo el resto conformando los datos de entrenamiento.

Selección del modelo

Supongamos que estamos tratando de resolver un problema real, mediante el uso de técnicas de aprendizaje automático. Recopilo información, genero un conjunto de datos etiquetado. Subdivido el conjunto de datos en el subconjunto de entrenamiento “Train set” y en el conjunto de pruebas “Test Set”. Ahora la pregunta es ¿qué modelo utilizo?

Dividir los datos solo en “Train Set” y “Test Set” no siempre es suficiente, principalmente cuando estamos haciendo distintos ejercicios entre varias funciones hipótesis de un determinado algoritmo de ML. Puede darse el caso de que el modelo ya haya visto los datos de prueba antes, mientras elige entre múltiples hipótesis.

Para resolver esto, lo que se haca es dividir nuestro conjunto inicial de datos en 3 subconjuntos:

• Utilizaremos el “Train Set” para entrenar los modelos.
• Utilizaremos el “Validation Set” para evaluar si se está produciendo Overfitting sobre el “Train Set”.
• Utilizaremos el “Test Set” para comprobar que el modelo que mejor se comporta para el “Train Set” y “Validation Set”, es capaz de generalizar bien tanto para ejemplos que no vio tanto en el “Train Set” y Validation Set”.

Resumen de selección de modelo

Como aplicaríamos este procedimiento en un caso real en donde deberé evaluar diferentes modelos, diferentes funciones hipótesis y ver cuáles de ellas se adaptan mejor a mi problema.

• Decido evaluar cinco funciones hipótesis (la primera en regresión lineal y el resto es regresión polinómica).
• Divido mi conjunto de datos en 3 subconjuntos (Train, Validation y Test).
• Entreno, es decir, buscar los parámetro w óptimos que minimicen mi función de error para el “Train Set” y calculo el error
• Calculo el error que hay respecto a mi subconjunto de entrenamiento,
• Elijo el modelo que tenga bajo los dos errores calculados anteriormente (supongamos que el modelo polinómico grado3, pintado).

• Vuelvo a evaluar el modelo seleccionado con el “Test Set”. Si este error es pequeño, entonces efectivamente ese modelo es el óptimo y resolví mi problema.