Respuestas Examen-Modulo IV

Edgar Adrian López González
ITAM

May 18, 2020

Pregunta 1

Podemos definir un programa de Machine Learning respecto a tres características. Menciona cuáles son

Respuesta:

- **Tarea:** Se puede definir como el objetivo o el propósito del programa. ***Ejemplo***: Pronosticar un precio o clasificar una serie de imágenes
- **Experiencia:** Es la *fuente* de posible datos (información) de donde el programa se basará para *aprender.****Ejemplo***: Una base de datos o el tomar una clase
- **Medida de desempeño:** Es la forma(métrica) con la cual se evaluará que tan bien o mal estamos realizando la *tarea.****Ejemplo***: La cantidad de veces que acetarmos en el tiro al blanco.

Pregunta 2

Describe qué es un método de aprendizaje supervisado

Respuesta:

Un método de aprendizaje supervisado es un programa de Machine learning que busca predecir o estimar un resultado basado en una serie de valores. Matemáticamente, se busca encontrar una función $\hat{f}$ que nos permite aproximar la relación f que existe entre una variable de interés Y y un conjunto de variables $X_1, X_2, ..X_n$.

Pregunta 3

Describre qué es un método de aprendizaje no supervisado

Respuesta:

Un método de aprendizaje no supervisado es un programa de Machine learning que busca encontrar las relaciones o posible estructura que describe a un conjunto de variables $X_1, ...X_n$. Algunos ejemplos pueden ser determinar agrupaciones que puedan estar presenten en los datos, o bien, la función de distribución de probabilidad de estos.

Pregunta 4

Define, en términos de una función, qué es un problema de regresión

Respuesta:
Un problema de regresión busca encontrar/estimar/aprender una función f que mapea un vector de variables $(X_1, .., X_n) \in \mathbb{R}^n$ a un valor numérico $Y \in \mathbb{R}$. Es decir, se busca encontrar f tal que:

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}$$
$$Y = f(X)$$

Pregunta 5

Define, en términos de una función, qué es un problema de clasificación

Respuesta:

Un problema de clasificación busca encontrar/estimar/aprender una función f que mapea un vector de variables $(X_1, .., X_n) \in \mathbb{R}^n$ a un valor discreto o categorico $Y \in \mathbb{C}_i$. Es decir, se busca encontrar f tal que:

$$f : \mathbb{R}^n \longrightarrow \{\mathbb{C}_i\}_{i=1}^N$$
$$Y = f(X)$$

Pregunta 6

Considerando cada uno de los siguientes pares de modelos, encierra en un circulo el que tenga mayor complejidad. Argumenta tu elección

Árbol de Decisión —— Random Forest
Regresión Lineal —— Regresión Ridge

Respuesta:
EL par de modelos que representa una mayor complejidad son el **Árbol de Decisión y Random forest**. Recordemos que en general, un modelo más complejo es aquel cuya estructura es más flexible a aprender las relaciones entre los datos, lo cual se traduce la

mayoria de las veces en un número grande de parametros a estimar, esto provoca un sesgo bajo pero una alta varianza. En general los modelos basados en árboles tienen un mayor número de parámetros a estimar por lo que sufren de bajo sesgo pero alta varianza, siendo modelos más complejos que los lineales, pues estos ultimos al imponer una estructura lineal a la posible relación entre los datos y penalizar el número de parametros (regresión ridge) dan una estrucutra menos flexible teniendo asi modelos con baja variabilidad pero un posible alto sesgo.

Pregunta 7

Considera una base de datos $D = \{(x_n, t_n)|x_n \in \mathbb{R}^M, t_n \in \mathbb{R}\}_{n=1}^N$ ¿Cómo asumimos se distribuye $t_n|x_n$ en una regresión lineal?

Respuesta:

En un modelo de regresión lineal simple se asume que la distribucion condicional de t_n dado x_n sigue la siguiente ley de probabilidad:

$$t_n|x_n \sim N(w^T x_n, \beta^{-1})$$

En otras palabras, se asume una distribución Normal en los datos, con el predictor lineal $w^T x_n$ como la medida y con precisión constante(homocedasticidad) $\beta = \frac{1}{\sigma^2}$.
Cabe mencionar que suponiendo este modelo, la estimación de w esta dada por

$$w^* = (X^T X)^{-1} X^T t_n$$

Pregunta 8

Considera una base de datos $D = \{(x_n, t_n)|x_n \in \mathbb{R}^M, t_n \in \{0, 1\}\}_{n=1}^N$ ¿Cómo asumimos se distribuye $t_n|x_n$ en una regresión logistica?

Respuesta:

En un modelo de regresión logística se asume que la distribucion condicional de t_n dado x_n sigue la siguiente ley de probabilidad:

$$t_n|x_n \sim \text{Bernoulli}(\sigma(w^T x_n))$$

En otras palabras, se asume una distribución bernoulli en los datos, con probabilidad p dada por la función sigmoidal evaluada en el predictor lineal $(w^T x_n)$, es decir:

$$\sigma(w^T x_n) = \frac{1}{1 + e^{-w^T x_n}}$$

Cabe mencionar que suponiendo este modelo, la estimación de w no puede determinarse por una formula cerrada, por lo que deben usarse algoritmos de optimización numerica como gradiante en descenso, newton-raphson, L-BFGS O gradiane en descenso estocastico.

Pregunta 9

Define, en términos de funciones (modelos) y bases de datos, la diferencia entre Voting, Boosting y Bagging

Respuesta:

Los siguiente metodos son metodos de ensamble, es decir, se agregan/poderan los resultados de distintos modelos de machine learning para tener un mejor performance. Si suponemos que contamos con una base de datos $D = \{(x_n, t_n)|x_n \in \mathbb{R}^M, t_n \in \{0,1\}\}_{n=1}^N$ de N elementos.

- **Voting** En este método se generan B modelos de regresión o clasificación, $y_i(x|D)$ con $i \in \{1...B\}$, dependiendo el tipo de problema con la ***misma base de datos*** D. Finalmente se promedian las estimaciones(en problemas de regresión) o la mayoria de votos (problemas de clasificación).

$$y_{vot}(x) = \frac{1}{B}\sum_{i=1}^{B} y_i(x|D)$$

- **Bagging**
  En este método se generan B muestras con reemplazo del mismo tamaño que la base de datos D, es decir, de N elementos. Posteriormente se entrena un modelo $y(x|D_i^*)$ con $i \in \{1...B\}$ para cada muestra . Finalmente se promedian las estimaciones(en problemas de regresión) o la mayoria de votos (problemas de clasificación).

$$y_{bag}(x) = \frac{1}{B}\sum_{i=1}^{B} y^*(x|D_i^*)$$

- **Boosting**
  En este método se van generando modelos de manera secuencial sobre variaciones de la base de datos original D. Es decir, dado un modelo $y_i(x)$ se ajusta un nuevo modelo $y_j(x)$ sobre los residuales o desviaciones del modelo actual, esto implica ir modificando la base de datos D sobre la que se esta trabajando.Finalmente se promedian las estimaciones(en problemas de regresión):

$$y_{boos}(x) = \sum_{i=1}^{B} \lambda y^*(x|D_i^*)$$

  Donde λ es un coeficiente de aprendizaje (que tan rapido o lento va aprendieno el modelo)

Pregunta 10

Si condieramos una función de costos $L(t, y(x)) = (t - y(x))^2$, la esperanza de la función de costos estaría dada por:

$$\mathbb{E}[L] = \int \{(\mathbb{E}_D[y(x|D)] - \mathbb{E}_t[t|x])^2 + \mathbb{V}_D[y(x|D)] + \mathbb{V}_t[t|x]\} P(x) dx$$

Describe cada uno de los componentes de la ecuación anterior y especifíca qué está midiendo

Respuesta:

Recordemos que cuando uno busca ajustar un modelo a un conjunto de datos D, busca que los valores de la medida de desempeño P sean los más adecuados, es decir, si suponemos que nuestra medida de desempeño es $L(t, y(x)) = (t - y(x))^2$ lo que esperamos es que el modelo $y(x)$ entrenado dada una base de datos D , en promedio se equivoque, lo menos posible. Por lo que la ecuación anterior nos permite entender que factores influyen al momento de evalular el desempeño de un modelo.

El termino $(\mathbb{E}_D[y(x|D)] - \mathbb{E}_t[t|x])^2$ representa el *sesgo*, es decir, un error que mide las desviaciones que tiene las estimaciones, dada cualquier base de datos D con la que es entrenado el modelo (por eso la esperanza), respecto a los verdadera media del distribución condicional.

El termino $\mathbb{V}_D[y(x|D)]$ representa la *varianza*, es decir, un error que mide la sensibilidad de nuestras desviaciones ante cambios en la base de datos D con la que es entrenado el modelo. Idealmente la estimación de $y(x|D)$ no deberia de cambiar mucho ante distintas bases de datos.

El termino $\mathbb{V}_t[t|x]$ representa el *error irreducible*, es decir, un error que mide la desviación de los valores observados respecto a su verdadera media, dado el modelo no conocido $t_n|x_n$. Este error es irreducible porque no importa que tan bien estimamos $y(x|D)$, este error siempre estara presente.

Pregunta 11

Describe el propósito y uso de k-fold Cross-Validation

Respuesta:

Recordemos que k-fold cross-validation es un método de remuestreo, que divide nuestra base de datos en k subconjuntos del mismo tamaño. De manera iterativa, se entrena el modelo con k-1 subconjuntos dejando siempre 1 partición como conjunto de validación y sobre la que calcularemos la metrica de performance del modelo. Esto nos permite generar k metricas de performance que en promedio nos dara una mejor estimación del performance del modelo

propuesto ante datos no observados.

Por ejemplo, para un modelo de regresión, el método K-fold cross validation nos genera una mejor estimación del error de test,

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}(t_i - y(x_i))^2$$

Lo anterior nos permite detectar y reducir la posibilidad de *overfitting*, al considerar k estimaciones de $y(x|D*)$ en k bases de datos *distintas*.