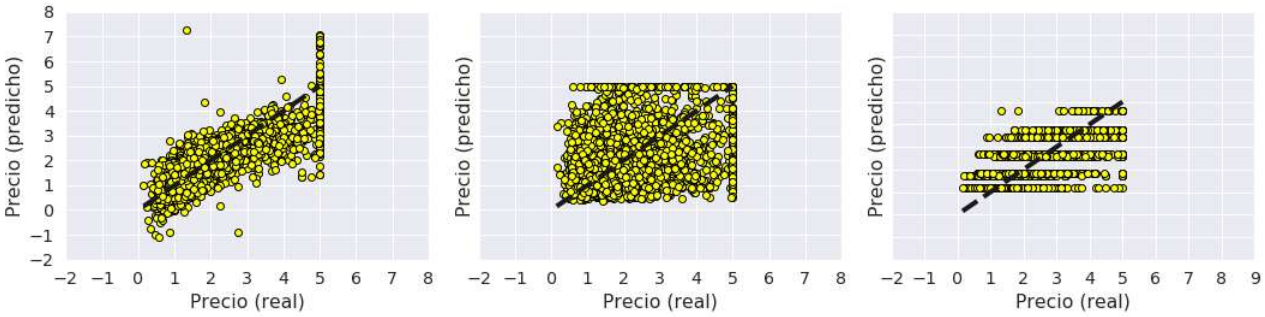
10. Regresión y Regularización

Ejercicio 10.1. Verdadero o Falso

1. El algoritmo de árboles de decisión no puede ser adaptado para un problema de regresión ya que la selección del mejor corte $< atributo, corte >$ es imposible cuando la variable a predecir es continua.
2. KNN para regresión con K = n (tamaño del conjunto de datos) devuelve una única región de decisión, para cada nueva instancia siempre se devolverá el mismo número y este número será el promedio de los datos de entrenamiento.
3. En árboles de decisión para regresión, los dos cambios principales son (a) utilizar, por ejemplo, la Varianza de lar región como medida de impureza de una región, (b) calcular el valor de una hoja como el promedio de las instancias que caen en esa región.
4. En árboles de decisión para regresión, si utilizamos MSE como medida de impureza, tomando como valor "real" para la región al promedio de las instancias, es equivalente a reducir la varianza.

Ejercicio 10.2. En la siguiente imagen puede verse el resultado de tres tipos de modelos distintos entrenados sobre los datos del dataset *California Housing Dataset*. Cada gráfico muestra el valor real y el valor predicho para instancias no vistas en entrenamiento (conjunto de validación). Aclaración: el precio de las propiedades en esta base de datos se mueve entre 0 y 5.

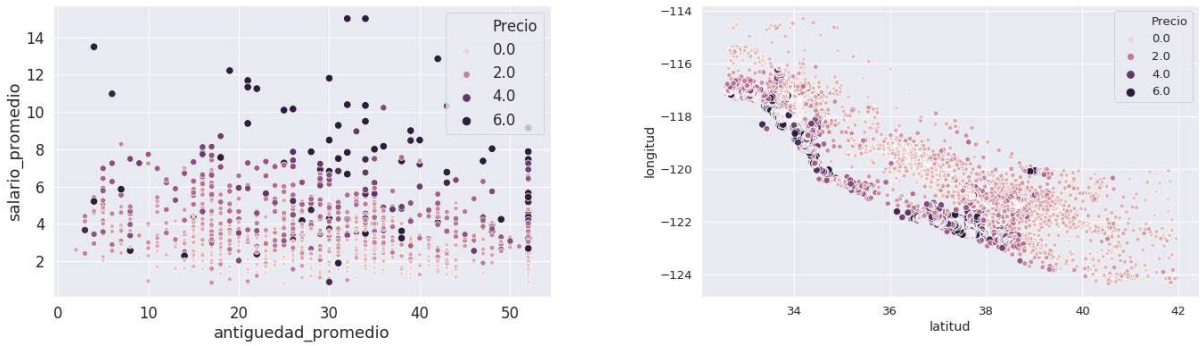


(a) ¿Qué se esperaría de estos gráficos si la regresión fuese perfecta?

(b) Determinar qué figura corresponde a una Regresión Lineal, cuál a un Decision Tree Regressor (binario, de profundidad máxima = 3) y cuál a un KNN Regressor (K = 1). Justifique.

(c) En el caso del árbol de decisión, ¿podemos decir algo del tipo de datos de los atributos (es decir, determinar si se trata de atributos numéricos o discretos)?

Ejercicio 10.3. El dataset *California Housing Dataset* [5] contiene una serie de instancias que representan zonas de California en las que interesa conocer el valor promedio de las viviendas. Cada instancia está representada por atributos tales como latitud, longitud, antigüedad promedio de las viviendas, y salario promedio de las personas en la zona, entre otros. A continuación se muestran dos gráficos generados a partir de estos datos:



[5] http://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html

Imagine que se ajusta una regresión lineal a dichas instancias. Dibuje una representación de cómo esperaría que dicha regresión impacte sobre cada uno de estos gráficos.

Ejercicio 10.4. Contestar las siguientes preguntas en el contexto de regresión, justificar:

1. ¿Cómo afecta la normalización de los datos para el modelo de regresión lineal sin regularización?
2. ¿Qué sucede al momento de aplicar regularización Ridge por ejemplo?
3. ¿Afecta al error MSE aplicar o no aplicar normalización a los atributos para el cálculo del error?
4. ¿Afecta la escala de Y (si cambiamos la salida y las etiquetas de km. a mts. por ejemplo) en las decisiones que toma el modelo si utilizamos MSE como métrica a optimizar?

Ejercicio 10.5. Gradient Descent

1. Escribir el pseudocódigo de la función "Descenso por gradiente", comentando brevemente qué se espera de cada argumento (junto a su tipo).
2. Comentar cuál es la función a minimizar en el caso de una regresión lineal.
3. Explicar cómo se obtiene el gradiente de la función a minimizar en el caso de una regresión lineal.
4. Escribir el pseudocódigo de *mini-batch gradient descent*.

Ejercicio 10.6.

A) Demuestre que tanto en ridge regression como en lasso regression cuando λ (i.e., el parámetro que controla la cantidad de regularización) tiende a infinito, las predicciones del modelo tienden a la media de la variable a predecir en el conjunto de entrenamiento.

B) Demuestre que si se penaliza también por el valor de w_0, las predicciones del modelo serán igual a 0.

C) Demuestre que la suma de los errores al cuadrado de entrenamiento en caso de penalizar la constante son mayores o iguales a los obtenidos en caso de no penalizarla.

Ejercicio 10.7. Pensando al error cuadrático medio (MSE) como una función de pérdida:

$$\mathrm{MSE}_{\mathrm{X,y}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{h}(x^{(i)}) - y^{(i)}\right)^2$$

Verdadero o Falso:

1. Esta métrica se define de esta manera para regresión lineal, para otros métodos (tal como árboles de regresión) hay que redefinir la fórmula anterior.
2. La función, vista como una función de pérdida, es siempre convexa.
3. Considerar las siguientes afirmaciones en el contexto de regresión lineal con pesos w:
   I) Se puede usar como función de pérdida viéndola como una función de X e y (los datos).
   II) Se puede usar como función de pérdida viéndola como una función de w.
   III) La función, vista como una función de pérdida, es siempre convexa.
   IV) Se busca minimizar esta función para X e y fijos en cada iteración de descenso de gradiente.
   V) En mini-batch gradient descent, en cada batch se minimiza una función de pérdida distinta.

Ejercicio 10.8. Verdadero o Falso

1. Entrenar una regresión lineal significa minimizar los pesos w.
2. Entrenar un modelo significa minimizar una función de costo y esta función depende de los pesos w.
3. Regularizar significa minimizar el MSE(w) manteniéndolo cerca de 0.
4. Regularizar significa minimizar el MSE(w) sumado a algún término de costo c(w) que habla de la magnitud de estos pesos.

5. Para un problema como el de predicción de casas, el w asociado a una variable como “distancia al obelisco” no depende de la unidad. Es decir, obtendremos el mismo w si medimos la distancia en km. o en mts.
6. Existen otras funciones de costo que permiten ajustar los w, por ejemplo, el error absoluto promedio (MAE).
7. La métrica MSE(w) penaliza w grandes.

Ejercicio 10.9. Resolver el notebook `notebook_descenso_gradiente.ipynb`.

Ejercicio 10.10. (opcional) **Huber Regression**
Consideremos la métrica de regresión conocida como Huber Loss δ definida como:

$$L_\delta(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{si } |y - \hat{y}| \leq \delta \\ \delta\left(|y - \hat{y}| - \frac{1}{2}\delta\right) & \text{si } |y - \hat{y}| > \delta \end{cases}$$

Esta función de perdida se puede utilizar para entrenar un modelo de regresión lineal, en cuyo caso se suele hablar de una *regresión de Huber*.

1. Graficar para un umbral fijo δ la pérdida de Huber, MSE y MAE en función del error $y - \hat{y}$.
2. Describe con tus palabras qué hace esta función de pérdida. ¿Cómo se comporta para errores pequeños y para errores grandes en relación con el umbral δ (fijo)?
3. Dado un conjunto de datos fijos, ¿cómo impacta en la métrica aumentar el valor de δ? ¿y disminuirlo? En ambos casos, ¿cómo cambia el ajuste del modelo?
4. Imagina que un conjunto de datos tiene algunos outliers. ¿Cómo crees que la regresión de Huber maneja estos outliers en comparación con la regresión de mínimos cuadrados (MSE) y la regresión con MAE? Teniendo esto en cuenta, ¿en que casos crees que sería útil usar este tipo de regresiones?