Mentoría: Coronavirus en Argentina y el mundo
Práctico 3: Introducción al Machine Learning.

Matías A. Bettera Marcat Lucas C. Didoné Daniel Peralta

December 14, 2020

1 Práctico Introducción al Machine Learning

Consigna:

Proponer un modelo de Aprendizaje Automático para estos datos.

Para ello deberan explorar y probar varios modelos, buscando las configuraciones que mejores resultados den. Tener en cuenta:

- Elección de la variable objetivo y features
- Selección de un modelo
- Ajuste de hiperparámetros
- Evaluación

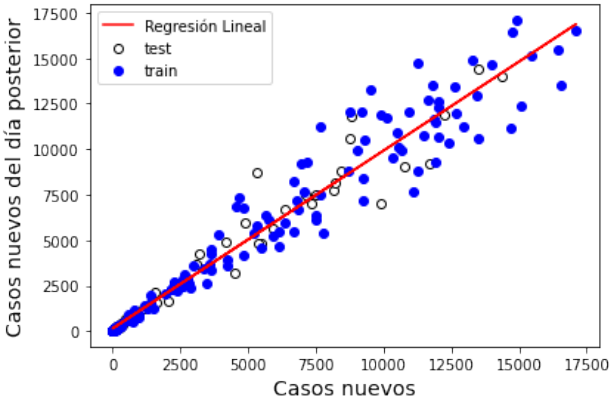
Veamos unos sencillos (e incompletos) **ejemplos**.

Variable objetivo: 'casos_nuevos': Puesto que muchas de las variables presentes en el dataset son **series temporales**, es importante tener en cuenta que para aplicar modelos de aprendizaje supervisado, es necesario realizar algunas adaptaciones previamente.

2 Regresión Lineal

Probaremos ajustar los puntos usando una recta.

```
Coefficients:
 [[0.97786561]]
Mean squared error
Train error: 1177344.95
Test error: 1042548.57
Coefficient of determination
Train: 0.95
Test: 0.94
```

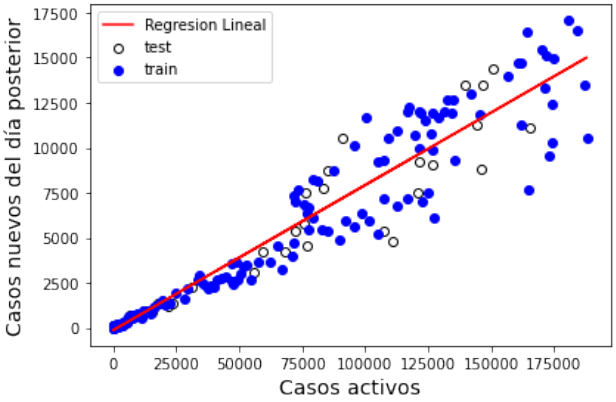


La Regresión lineal de casos nuevos versus casos nuevos del día anterior es muy buena, como se podía esperar, pero tiene una dispersión muy grande y que crece en el tiempo. Además no da mucha información, en el sentido de que no es posible advertir un cambio de signo en la curva de casos nuevos solamente con la predicción de un día. Además el hecho de que dependa del dato del día anterior hace que esté sujeto a variaciones propias de la dispersión de los datos y arrastradas por las irregularidades en la carga de datos.

Regresión casos activos versus casos nuevos

Regresión Lineal: Probaremos ajustar los puntos usando una recta.

```
Coefficients:
 [[0.08065208]]
Mean squared error
Train error: 1922454.78
Test error: 1974228.74
Coefficient of determination
Train: 0.92
Test: 0.90
```

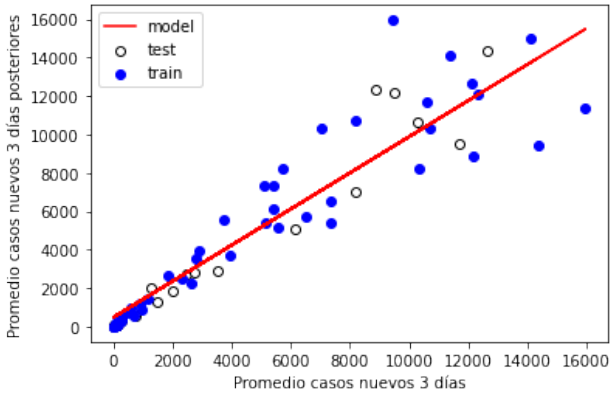


Aquí propusimos ajustar casos nuevos con casos activos, debido a que se supone que los activos en un período determinado son los que pueden contagiar y por lo tanto los contagiados nuevos deberían depender de los activos en tiempo anterior. Si bien sabemos que esto no es un modelo correcto ya que la dependencia es más compleja, puede ser un modelo que dé más información que el previo. Los datos siguen teniendo un buen ajuste lineal, si miramos los coeficientes determinación para Train y Test siguen siendo altos (>0.9), pero ahora se pueden distinguir dos regiones. Cuando los casos activos eran menos de 8000, se observa un régimen más lineal con poca dispersión de los datos, y a partir de que los casos activos superan los 8000, hay una gran dispersión de los datos que aunque puedan ser ajustados linealmente, claramente no se está reflejando su verdadera dependencia.

Es necesario remarcar que cuando fueron aumentando los casos, los sistemas de diagnósticos y de los reportes de casos nuevos han tenido cierto retraso, combinado con un aumento del índice de positividad. Esa puede ser una de las causas de aumento abrupto en la dispersión de los datos.

Regresión lineal promediando de a 3 días

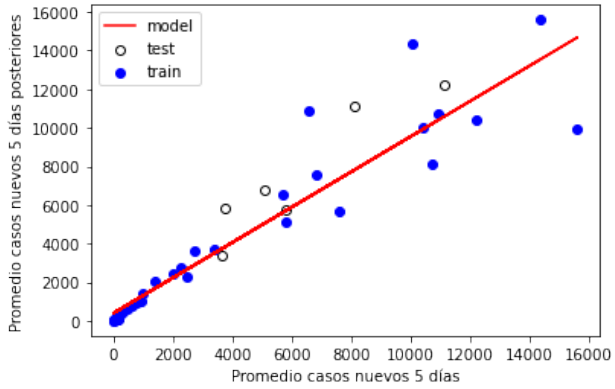
```
Coefficients:
 [[0.94036448]]
Mean squared error
Train error: 2891337.28
Test error: 1516197.78
Coefficient of determination
Train: 0.87
Test: 0.93
```



Regresión lineal promediando de a 5 días

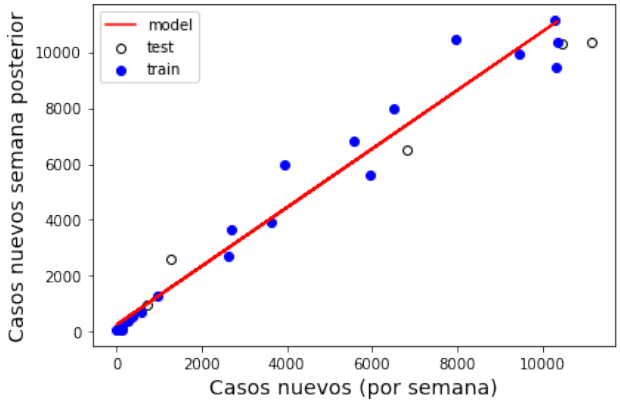
```
Coefficients:
 [[0.9152149]]
Mean squared error
Train error: 2375373.88
Test error: 2365507.79
Coefficient of determination
Train: 0.89
```

```
Test: 0.87
```



Regresión lineal promediando de a 7 días

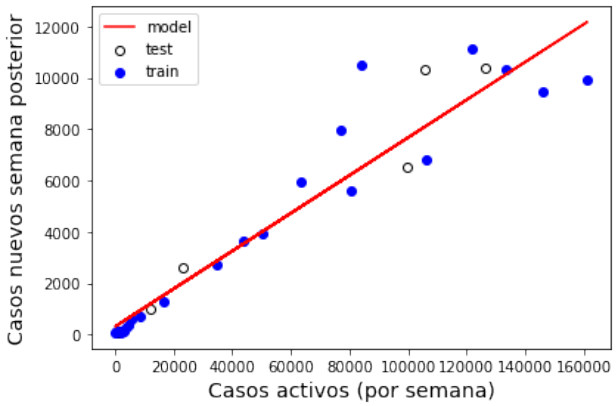
```
Coefficients:
 [[1.0470609]]
Mean squared error
Train error: 530077.55
Test error: 852513.88
Coefficient of determination
Train: 0.97
Test: 0.95
```



Conclusiones de promediar ventanas temporales: Encontramos un grado de correlación muy similar al promediar en ventanas de 3, 6 y 7 días, comparado con los datos sin promediar. En el caso de 7 días se puede observar que se reduce bastante la dispersión de los datos alrededor de la regresión, esto es positivo de cara a construir un modelo predictivo, pero hay que tener en cuenta que la información es un promedio semanal.

Casos activos versus casos nuevos

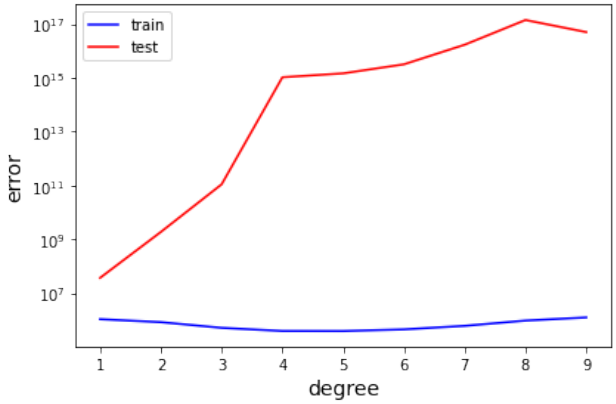
```
Coefficients:
 [[0.07348644]]
Mean squared error
Train error: 1507043.44
Test error: 1193003.15
Coefficient of determination
Train: 0.91
Test: 0.93
```



En este caso se observa que se mejora el coeficiente de determinación, gracias a utilizar promedios semanales. Aún así, la mejora se puede considerar marginal, y es importante notar que la dispersión de los datos sigue siendo alta. Este modelo puede ser un buen esquema para estimar cotas de nuevos casos semanales.

3 Regresión polinomial

degree	train err	test err
1	1.08e+06	3.70e+07
2	8.36e+05	1.88e+09
3	5.14e+05	1.11e+11
4	3.93e+05	1.06e+15
5	3.90e+05	1.48e+15
6	4.50e+05	3.20e+15
7	6.09e+05	1.71e+16
8	9.56e+05	1.42e+17
9	1.25e+06	5.05e+16

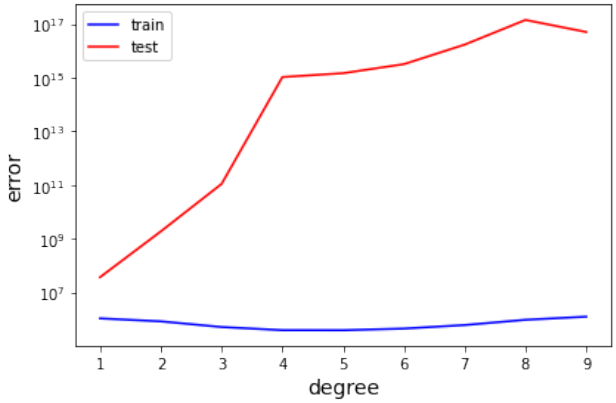


Por último, en los que se refiere a regresiones de datos temporales, probamos distintos grados polinomiales para 7 varialbes $('casos_nuevos','casos_activos','t_days','tests_nuevos','tests_total','muertes_nuevos'), conelobjetivodepredecircasosnuevos$

4 Experimentación con el dataset: Covid19Casos

Probamos el dataset provisto por el Ministerio de Salud. Este dataset contiene las fechas de apertura de legajo, de diagnóstico, de deceso, etc., de cada caso y del reporte en que se haya notificado a la prensa. Por ello es posible reconstruir una curva de datos de casos nuevos y casos totales en función del tiempo más realistas. Desafortunadamente no contamos en este dataset con los datos de altas, por lo que no se puede estimar los casos activos.

```
Coefficients:
 [[0.92972894]]
Mean squared error
Train error: 3686790.94
Test error: 2870772.57
Coefficient of determination
Train: 0.85
Test: 0.88
```



5 Clasificación

Tratamiento por género y rango/edad: Como se trata de **variables categóricas**, en la cuál no hay una relación de orden entre las categorías, necesitamos aplicar un **One Hot Encoding**

Selección de features y variable objetivo: variable objetivo = 'provincia'.

Logistic Regression

```
Accuracy of Logistic regression classifier on training set: 0.49
Accuracy of Logistic regression classifier on test set: 0.51
```