REGRESIÓN LINEAL

Aprendizaje Automático

CEIoT - FIUBA

Dr. Ing. Facundo Adrián Lucianna

LO QUE VIMOS LA CLASE ANTERIOR…

NUMPY

Cuando trabajamos en aplicaciones de Aprendizaje Automático, trabajamos con muchos tipos de datos y en grandes cantidades.

Los conjuntos de datos pueden provenir de una amplia gama de fuentes y formatos, como colecciones de documentos, imágenes, clips de sonido, mediciones numéricas, entre otros. A pesar de esta aparente heterogeneidad, nos ayudará pensar en todos los datos fundamentalmente como arrays de números.

El almacenamiento y manipulación eficientes de arrays numéricos son absolutamente fundamentales en el proceso de hacer Aprendizaje Automático.

Esto lo podemos hacer usando **NumPy**.

NumPy nos ofrece los arrays que puede almacenar y operar datos de manera eficiente.

- Una forma que logra ser eficiente es que los arrays sus elementos tienen que ser del mismo tipo.
- Los arrays son **mutables** por defecto, pero se puede cambiar este comportamiento.

ESTRUCTURAS DE DATOS DE PANDAS

A un nivel muy básico, los objetos de **Pandas** son una versión mejorada de las estructuras de NumPy en los cuales las filas y columnas se identifican con etiquetas.

Tiene dos estructuras fundamentales:

- Series: Una serie de Pandas es un array uni-dimensional de datos indexados.
- DataFrame: Es análogo a un array de dos dimensiones con índices de filas y nombres de columnas. Un DataFrame se forma con una Serie para cada columna.

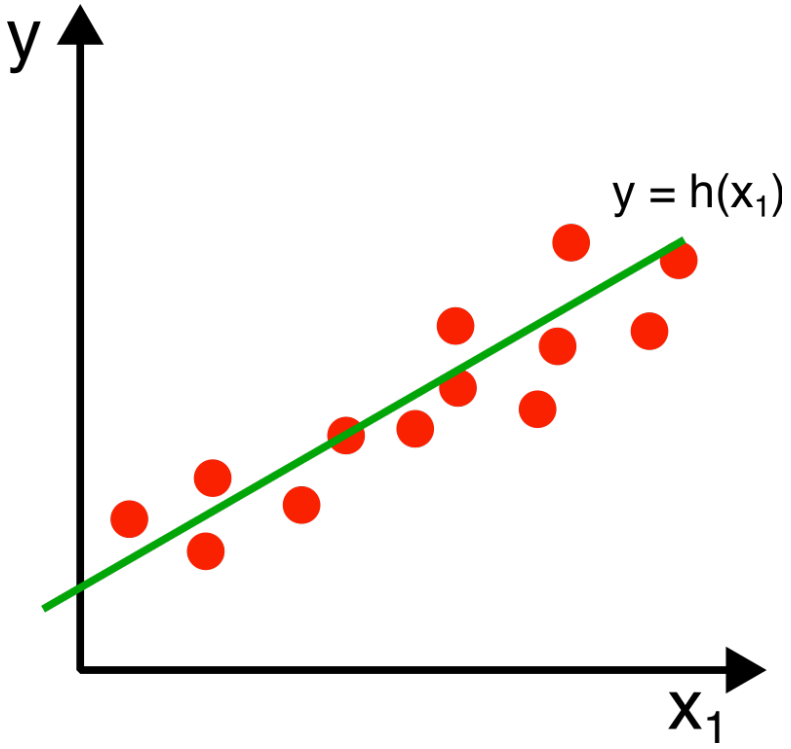
REGRESIÓN

REGRESIÓN

Si tenemos un problema donde el target **y** es una *variable numerica*, se llama un **problema de regresión**.

Se centra en estudiar las relaciones entre una variable dependiente de una o más variables independientes.

Es importante notar que, en Aprendizaje Automático, cuando buscamos una h(X) estamos armando un modelo puramente empírico. Es decir, nos basamos 100% en los datos medidos. En contraste con los modelos basados en propiedades fundamentales.



REGRESIÓN LINEAL

REGRESIÓN LINEAL

El modelo de regresión lineal más simple es el que involucra una combinación lineal de las variables de entradas:

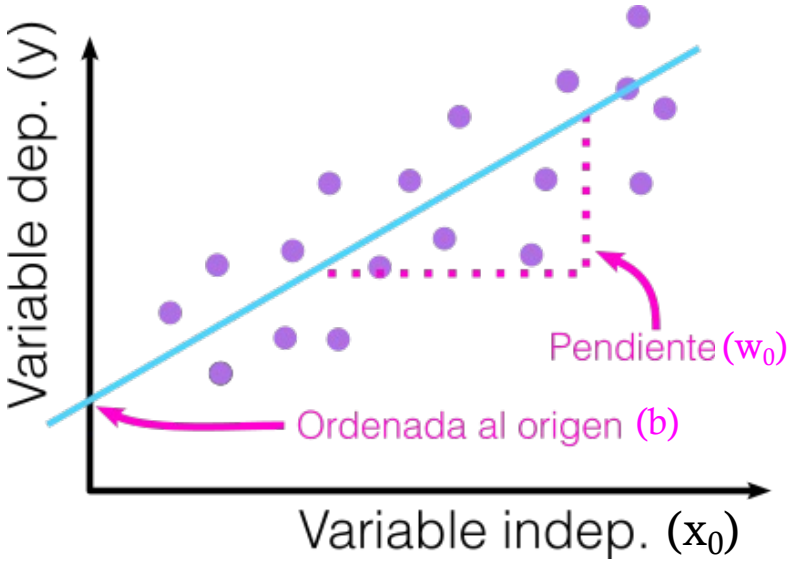
$$\hat{y} = h(X) = b + w_0 x_0 + \cdots + w_d x_d$$

- $X = (x_0, x_1, \ldots, x_d)$ — Son los *features* de nuestras observaciones. Son todas variables numéricas
- $b, w_0, \ldots, w_d$ — Son los coeficientes del modelo. Son números reales. Cuanto más cerca de cero, la variable dependiente depende menos del *feature* que multiplica.
- $\hat{y}$ — Es la predicción del modelo. Es con quien comparamos con el *Label* de la observación

REGRESIÓN LINEAL

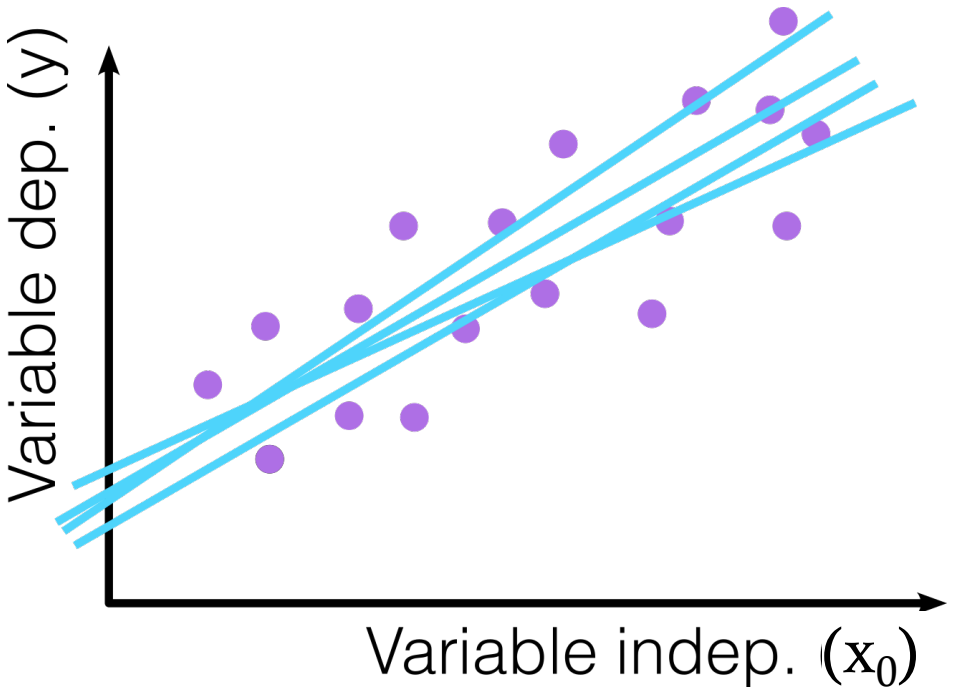
Vamos al caso más sencillo, la regresión lineal de una sola variable independiente:

$$\hat{y} = h(X) = b + w_0 x_0$$



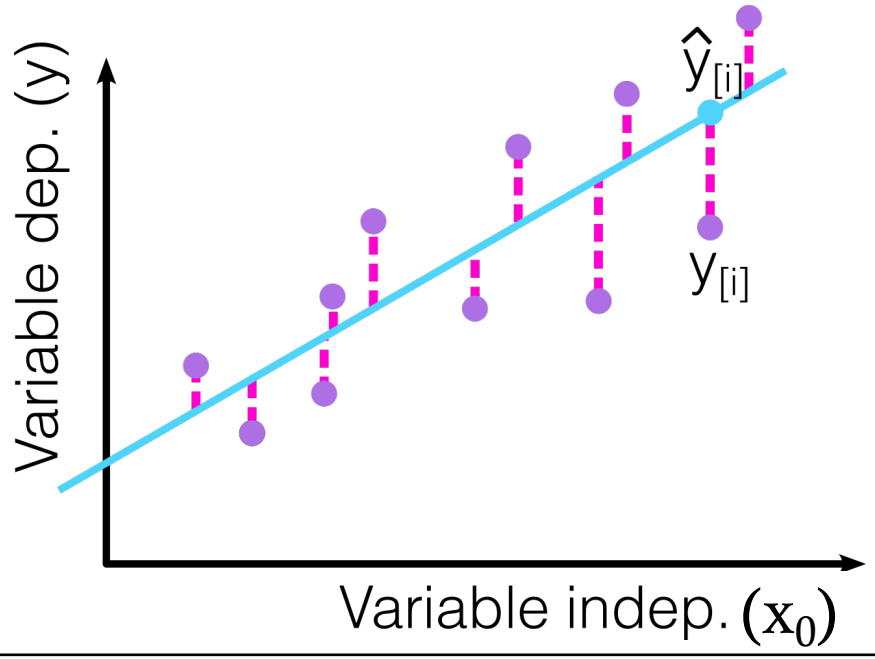
REGRESIÓN LINEAL

¿Ahora cuál recta?



REGRESIÓN LINEAL

Para encontrarla, medimos la distancia entre la recta y cada punto, que llamamos **residuos**.

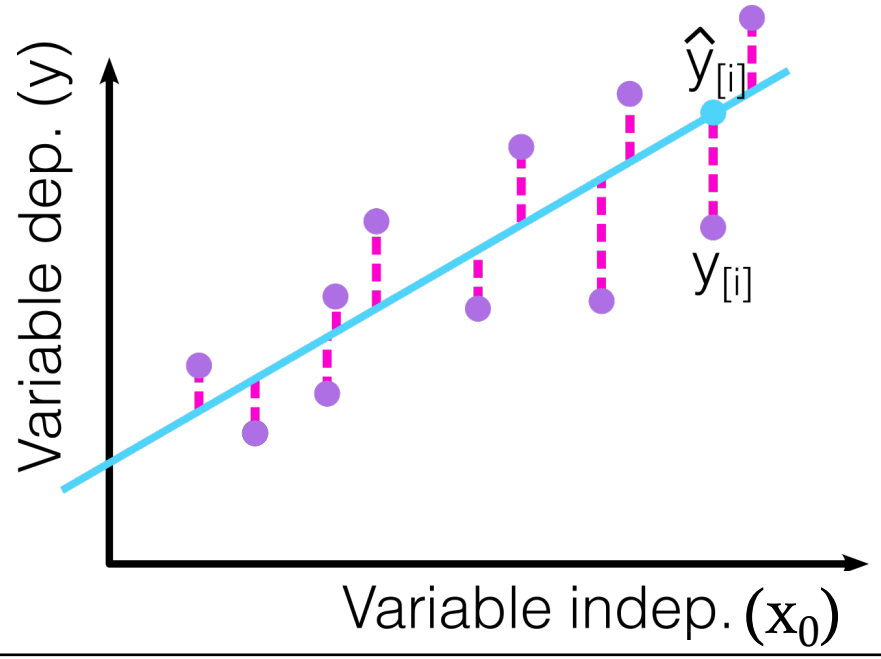


$$e_{[i]} = y_{[i]} - \hat{y}_{[i]}$$

$$y_{[i]} = b + w_0 x_{0[i]} + e_{[i]}$$

REGRESIÓN LINEAL

Buscamos minimizar el valor de los residuos. Para lograr esto, lo hacemos minimizando la suma de los cuadrados de los **residuos**.



$$S_R = \sum_{i=0}^{N-1} \left(e_{[i]}\right)^2 = \sum_{i=0}^{N-1} \left(y_{[i]} - b - w_0 x_{0[i]}\right)^2$$

$$\min(S_R) = \min\left(\sum_{i=0}^{N-1} \left(e_{[i]}\right)^2\right)$$

Para minimizar, solo podemos tocar los coeficientes. Lo que hacemos es ir por el **gradiente**.

$$\frac{\partial S_R}{\partial b} = 0 \qquad \frac{\partial S_R}{\partial w_0} = 0$$

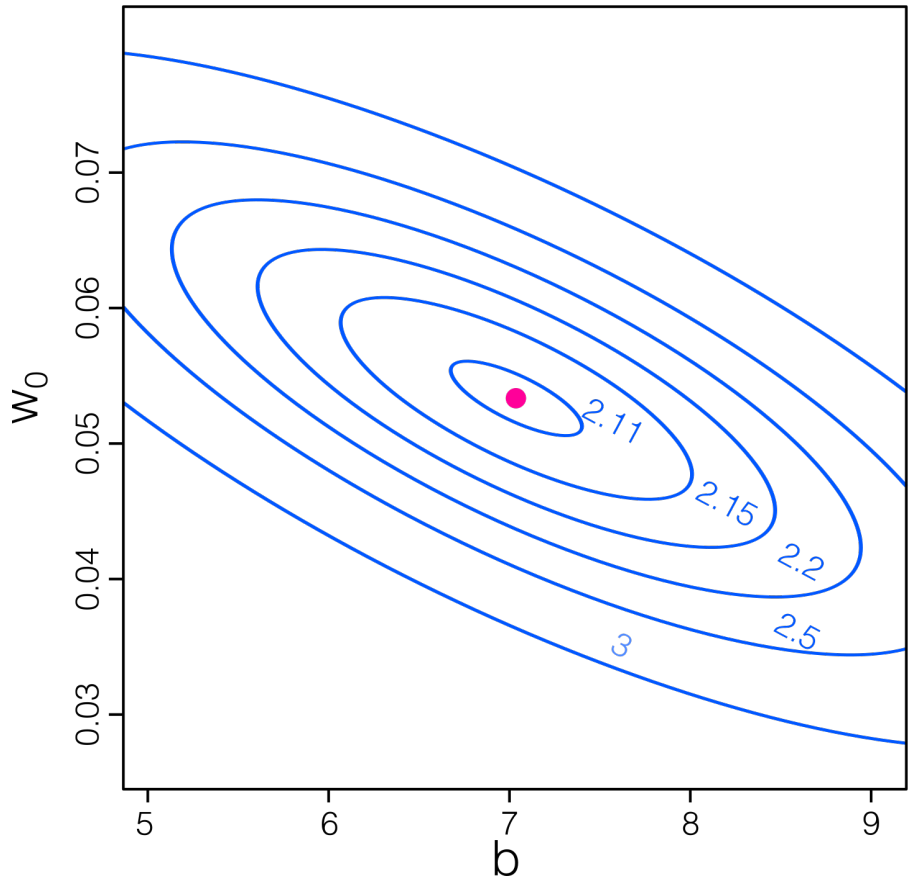
REGRESIÓN LINEAL

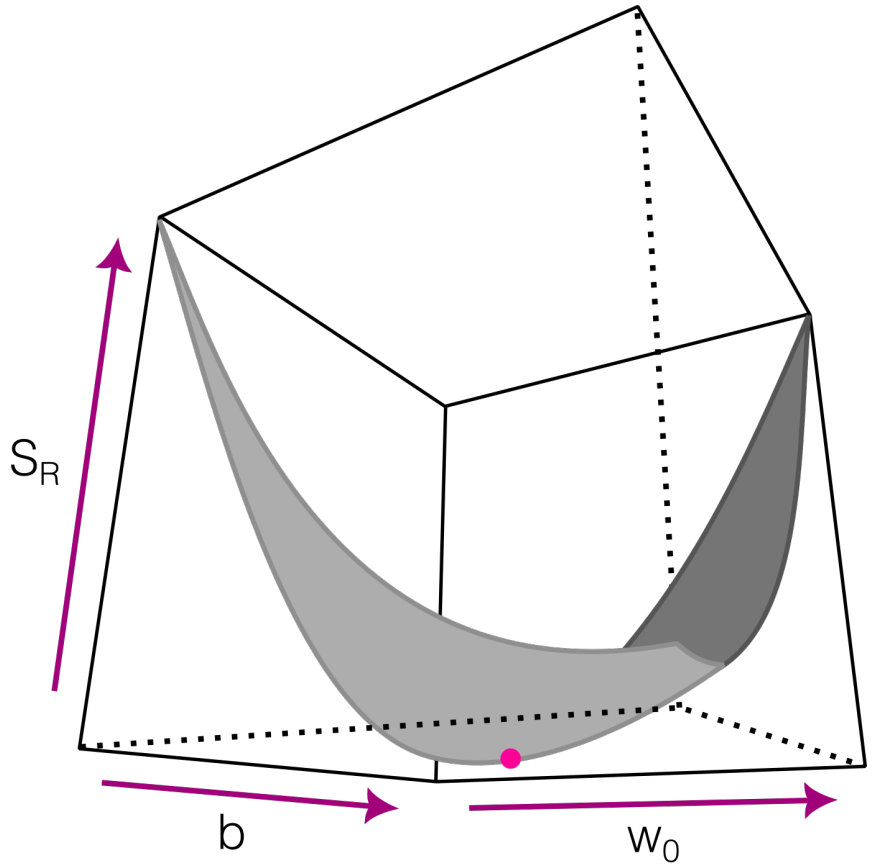
S_R en regresión lineal es siempre convexa, es decir que siempre tiene un solo mínimo. En su forma tradicional,

$$\frac{\partial S_R}{\partial b} = 0 \qquad \frac{\partial S_R}{\partial w_0} = 0$$

Si planteamos las derivadas, obtenemos un sistema de ecuación, llamado ecuaciones normales. Si se resuelve este sistema se encuentra la solución.

REGRESIÓN LINEAL





REGRESIÓN LINEAL

Suposiciones

Las suposiciones que usamos para poder aplicar regresión lineal son:

- **Relación lineal:** Lógicamente, y esto muchas veces al aplicar el modelo buscamos validar.
- **Features independientes:** Los features de entrada de la regresión deben ser independientes entre sí.
- **Homocedasticidad:** Es decir, la "nube" se mantiene igual en toda parte de la recta.
- **Errores independientes:** Los errores entre sí no están correlacionados.

MÉTRICAS DE EVALUACIÓN

MÉTRICAS DE EVALUACIÓN

Cuando se arma un modelo, al dataset lo separamos en una parte usada para entrenar el modelo y la parte de evaluación. El conjunto de datos de evaluación se utiliza para evaluar qué tan bien se entrenó el algoritmo con el conjunto de datos de entrenamiento.

¿Pero cómo evaluamos?

- **El coeficiente de determinación (R^2).** El coeficiente determina la calidad del modelo para replicar los resultados, y la proporción de variación de los resultados que puede explicarse por el modelo.

  Aunque no siempre es el mejor caso. Podemos usar métricas más generales, métricas que midan error de variables numéricas que se pueda aplicar también a otros tipos de casos, como por ejemplo forecasting en series de tiempo.

MÉTRICAS DE EVALUACIÓN

Error absoluto medio (MAE)

El **error absoluto medio (MAE)** es el cálculo del valor absoluto del residuo para cada punto de datos, para que los residuos negativos y positivos no se cancelen. Luego tomamos el promedio de todos estos residuos.

$$MAE = \frac{1}{N}\sum_{i=0}^{N-1} |y_{[i]} - \hat{y}_{[i]}|$$

Debido a que utilizamos el valor absoluto del residuo, MAE no indica si el modelo sobreestima o subestima.

MÉTRICAS DE EVALUACIÓN

Error cuadrático medio (MSE)

El **error cuadrático medio (MSE)** es similar al **MAE**, pero ahora calculamos el cuadrado de los residuos. Esto es similar a lo que se usamos para encontrar los coeficientes.

$$MSE = \frac{1}{N}\sum_{i=0}^{N-1}\left(y_{[i]} - \hat{y}_{[i]}\right)^2$$

MSE siempre es mayor a **MAE**. Un detalle importante son aquellos residuos grandes (**outliers**), en esta métrica aporta más que en **MAE**. En **MAE** el aporte es proporcional al valor del residuo, pero aquí es cuadráticamente más grande.

MÉTRICAS DE EVALUACIÓN

Raíz cuadrada del error cuadrático medio (RMSE)

Si al **MSE** le calculamos la raíz, tendremos una métrica llamada **RMSE** que tiene la misma unidad de la salida original, donde el **MSE** no. El **RMSE** es análogo al desvío estándar.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1}\left(y_{[i]} - \hat{y}_{[i]}\right)^2}$$

MÉTRICAS DE EVALUACIÓN

Outliers

Valores atípicos es una constante de discusión. *¿Se incluyen o no?*

La respuesta dependerá del problema en particular, de los datos disponibles y las consecuencias que hay si se consideran o no.

Si quiero tenerlo en cuenta a la hora de comparar modelos, me va a convenir usar MSE, en cambio sí quiero reducir su importancia, puedo usar MAE.

Ambas son métricas de error viables, pero describen diferentes matices sobre los errores de predicción.

MÉTRICAS DE EVALUACIÓN

Error absoluto porcentual medio (MAPE)

El **error absoluto porcentual medio (MAPE)** es el cálculo del error **MAE,** pero escalado al verdadero valor, por lo que el resultado es porcentual

$$MAPE = \frac{100\%}{N} \sum_{i=0}^{N-1} \left| \frac{y_{[i]} - \hat{y}_{[i]}}{y_{[i]}} \right|$$

No es una métrica buena porque es susceptible a errores numéricos. No puede calcularse cuando $y_{[i]}$ vale cero. Y además tiene sesgo para cuando la predicción subestima:

$n = 1 \quad \hat{y} = 10 \quad y = 20$
$MAPE = 50\%$

$n = 1 \quad \hat{y} = 20 \quad y = 10$
$MAPE = 100\%$

TRATAMIENTO DE VARIABLES

TRATAMIENTO DE VARIABLES

Normalización o estandarización

En la regresión lineal, tenemos la multiplicación de coeficientes por nuestras entradas:

$$\hat{y} = b + w_0 x_0 + w_1 x_1$$

Los coeficientes nos dan un **valor de importancia de las entradas**. Pero esto si todas las entradas están en la misma escala.

Si la variable x_0 está en rango de [1000, 3000] y x_1 en [-1, 1], los valores de w_0 y w_1 van a ser de diferentes escalas, y por consiguiente no comparables.

Además, aunque la regresión lineal no presenta problemas de escalas, valores muy diferentes nos pueden introducir **errores numéricos**. Otro tipo de regresiones o clasificadores si o si necesitan escalas, por lo que debemos normalizar o estandarizar.

TRATAMIENTO DE VARIABLES

Normalización o estandarización

Una forma de **normalizar** es hacer que los valores estén entre 0 y 1, tomando el máximo y el mínimo.

$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Esta fórmula se usa cuando la distribución de datos no es normal. Pero cuando nuestros datos tienen una distribución normal, como la suposición en regresión lineal, se aplica **estandarización**:

$$\tilde{x} = \frac{x - \text{mean}(x)}{\text{std}(x)}$$

Donde ahora hacemos que la distribución tenga medio cero y desvío estándar uno.

Con estos escalados, ahora los parámetros tendrán sentido entre sí.

TRATAMIENTO DE VARIABLES

Variables Dummies

Como vimos, regresión utiliza variables numéricas para predecir un valor. ¿Como podemos hacer para usar variables categóricas?

Para poder usarlos, debemos transformarlos en numéricas mediante alguna codificación.

Cuando tenemos variables categóricas ordinales, podemos asociar un número.

Por ejemplo, si tenemos que usar casos como : *"me gusta mucho", "me gusta poco", "neutral", "no me gusta poco"…*

Se puede usar números enteros, teniendo en cuenta el orden para darle importancia. Estará en la creatividad de quien lo hace para determinar si las distancias son equidistantes o no.

TRATAMIENTO DE VARIABLES

Variables Dummies

Ahora si tenemos casos nominales no podemos asociar números, porque al hacerlo, establecemos un orden.

Para este tipo de variable existe **one-hot encoding**.

Al usar esta codificación, creamos nuevos atributos de acuerdo con la cantidad de clases presentes en la variable categórica, es decir, si hay **n** número de categorías, se crearán **n** nuevos atributos. Estos atributos creados se denominan ***variables dummies***.

TRATAMIENTO DE VARIABLES

Variables Dummies

Peso	Altura	País
80	180	Argentina
83	177	Chile
75	169	Chile
68	155	Argentina

TRATAMIENTO DE VARIABLES

Variables Dummies

Peso	Altura	País	arg	chile
80	180	Argentina	1	0
83	177	Chile	0	1
75	169	Chile	0	1
68	155	Argentina	1	0

TRATAMIENTO DE VARIABLES

Variables Dummies

Peso	Altura	arg	chile
80	180	1	0
83	177	0	1
75	169	0	1
68	155	1	0

TRATAMIENTO DE VARIABLES

Variables Dummies

Como vimos, **one-hot enconding** nos genera un nuevo atributo por categoría, pero esto nos genera *una trampa*

Si vemos el ejemplo, las dos variables que estamos usando están 100% correlacionadas entre sí:

$$x_{chile} = 1 - x_{arg}$$

$$\hat{y} = b + w_0 x_{peso} + w_1 x_{altura} + w_2 x_{arg} + w_3 x_{chile}$$

$$\hat{y} = b + w_0 x_{peso} + w_1 x_{altura} + w_2 x_{arg} + w_3 (1 - x_{arg})$$

$$\hat{y} = (b + w_3) + w_0 x_{peso} + w_1 x_{altura} + (w_2 - w_3) x_{arg}$$

Para solucionar esto es quitar siempre quitar una columna para romper la trampa.

TRATAMIENTO DE VARIABLES

Variables Dummies

Peso	Altura	arg
80	180	1
83	177	0
75	169	0
68	155	1

REGRESIÓN LINEAL

Vamos a practicar un poco…

REGRESIÓN DE FUNCIONES BASES

REGRESIÓN DE FUNCIONES BASES

¿Qué pasa si queremos ver relaciones no lineales?

Un truco es adaptar la regresión lineal a relaciones no lineales es transformar los datos usando **funciones bases**.

La idea es tomar un modelo linear multidimensional:

$$\hat{y} = h(X) = b + w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$$

Y construir a x_1, x_2, …, de una entrada de una sola dimensión x_0.

Por ejemplo, si hacemos $f_n(x) = x^n$, nuestro modelo se transforma en una **regresión polinómica**:

$$\hat{y} = h(X) = b + w_0 f_1(x_0) + w_1 f_2(x_0) + w_2 f_3(x_0) + \cdots + w_{d-1} f_d(x_0)$$

$$\hat{y} = h(X) = b + w_0 x_0 + w_1 x_0^2 + w_2 x_0^3 + \cdots + w_{d-1} x_0^d$$

No solo podemos usar funciones polinómicas, sino que cualquier otra función.

REGRESIÓN LASSO Y RIDGE

REGRESIÓN DE RIDGE Y LASSO

Como hacemos para saber el aporte de cada atributo

Cuando se trata de entrenar modelos, nos podemos encontrar como problemas de sobreajuste, podemos implementar algún método de regularización.

Con estos métodos de regularización podemos ajustar un modelo que contenga todos los atributos utilizando una técnica que restrinja o regularice las estimaciones de los coeficientes o, de manera equivalente, que reduzca las estimaciones de los coeficientes hacia cero.

Puede que no sea inmediatamente obvio por qué tal restricción debería mejorar el ajuste, pero resulta que **reducir las estimaciones de los coeficientes** puede **reducir significativamente su varianza**.

Las dos técnicas más conocidas para reducir los coeficientes de regresión a cero son la regresión de **Ridge** y la de **Lasso**.

REGRESIÓN DE RIDGE

En la regresión lineal, vimos que se buscaba los coeficientes que minimizaban la suma de los residuos al cuadrado. La regresión de Ridge es muy similar, pero excepto que los coeficientes se estiman minimizando una cantidad ligeramente diferente:

$$\sum_{i=0}^{N-1} \left(y_{[i]} - b - w_0 x_{0[i]} - \cdots - w_d x_{d[i]} \right)^2 + \alpha \sum_{j=0}^{d} w_j^2$$

Donde α es un hiper-parámetro de ajuste.

REGRESIÓN DE RIDGE

En esta regresión, se busca los coeficientes que minimizan los residuos. Sin embargo, el segundo termino:

$$\alpha \sum_{j=0}^{d} w_j^2$$

Llamado el termino de penalización por encogimiento. Este es pequeño cuando los coeficientes están cerca de cero.

α funciona de control al impacto relativo de ambos términos. Cuando $\alpha = 0$, es una regresión normal. En cambio, si α crece, el impacto de termino de penalización crece, y los coeficientes se acercan a cero.

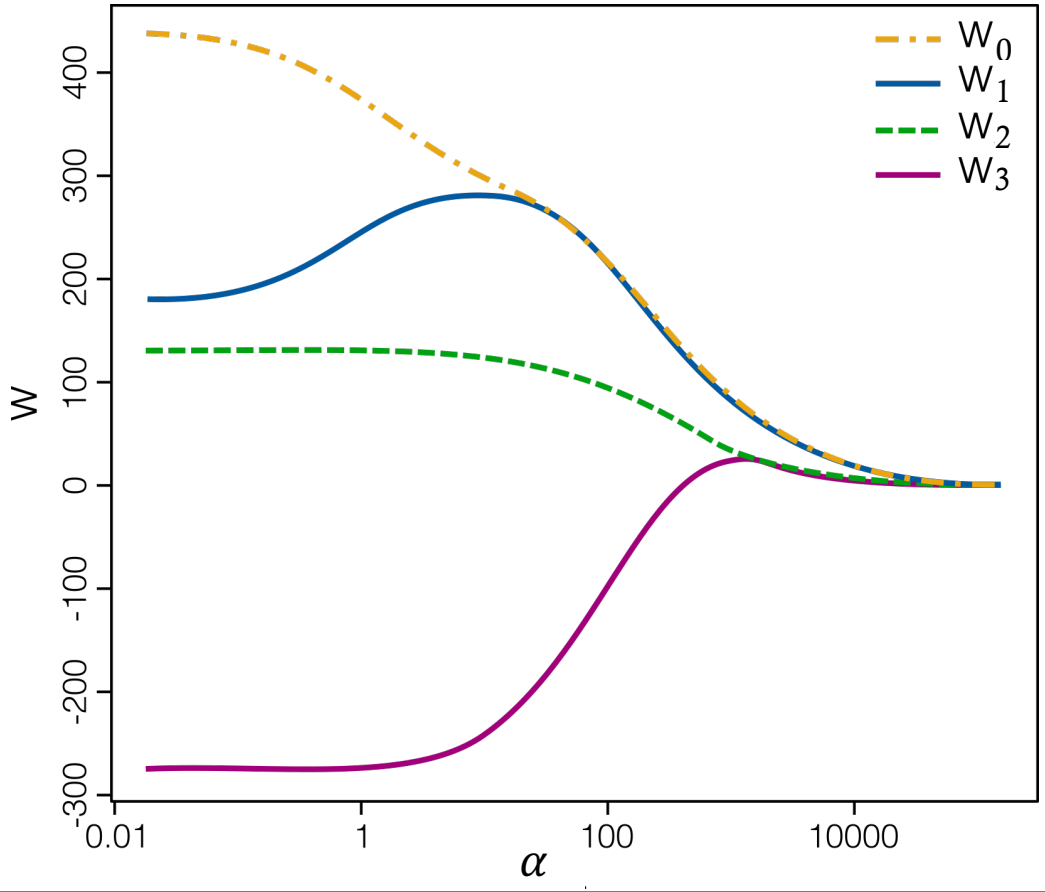
La **regresión de Ridge** genera un set de coeficientes por cada valor de α. Seleccionar un valor de α es difícil, y se deben usar métodos de validación cruzada.

REGRESIÓN DE RIDGE

Notese que la penalización no toca la ordenada al origen b. Si α es ∞, todos los coeficientes son cero y b nos queda:

$$b = \bar{y} = \frac{1}{N} \sum_{i=0}^{N-1} y_{[i]}$$

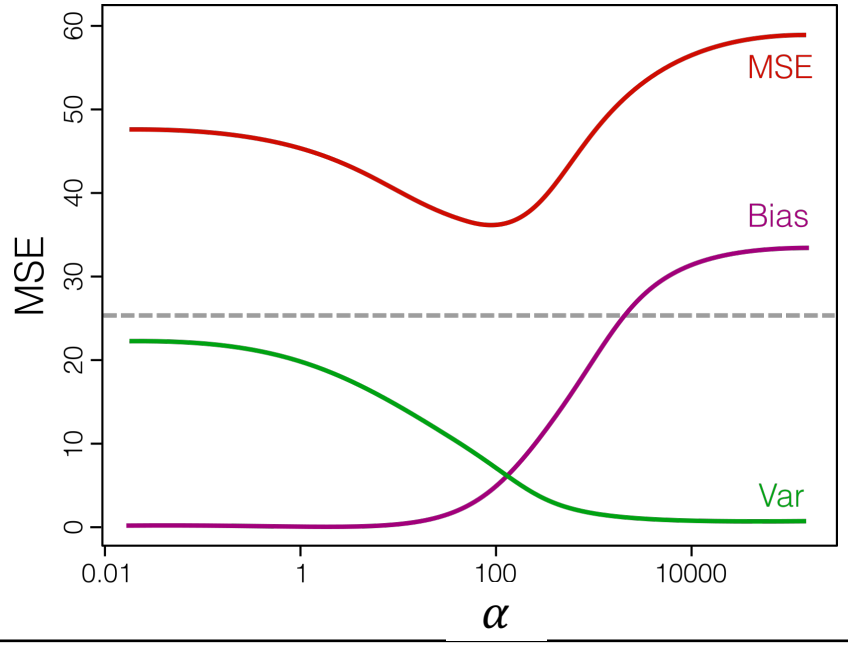
REGRESIÓN DE RIDGE



REGRESIÓN DE RIDGE

¿Para qué nos sirve?

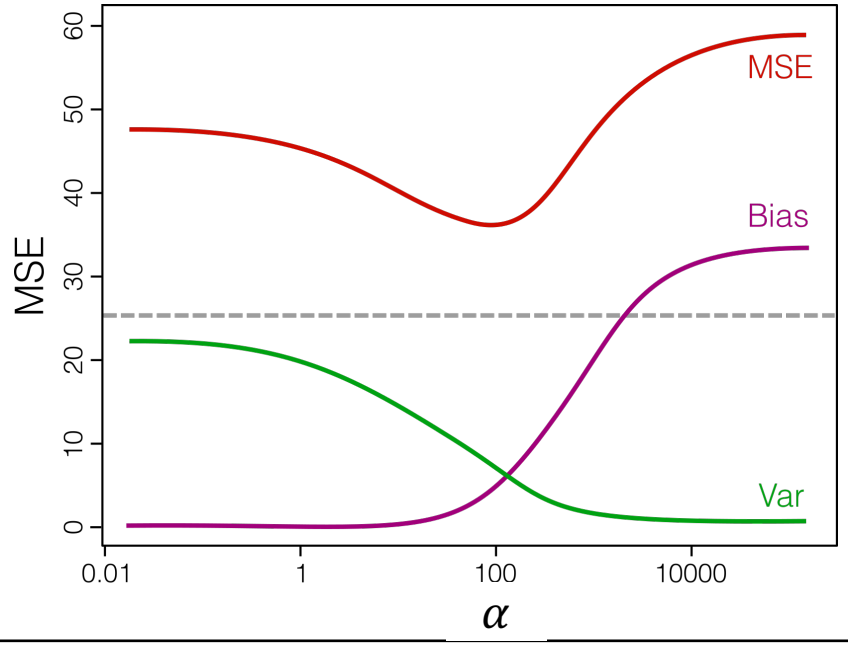
La ventaja de la regresión de Ridge sobre la regresión lineal de mínimos cuadrados tiene su origen en el equilibrio entre sesgo y varianza. A medida que α aumenta, la **flexibilidad del ajuste disminuye**, lo que lleva a una **menor varianza**, pero a un **mayor sesgo**.



REGRESIÓN DE RIDGE

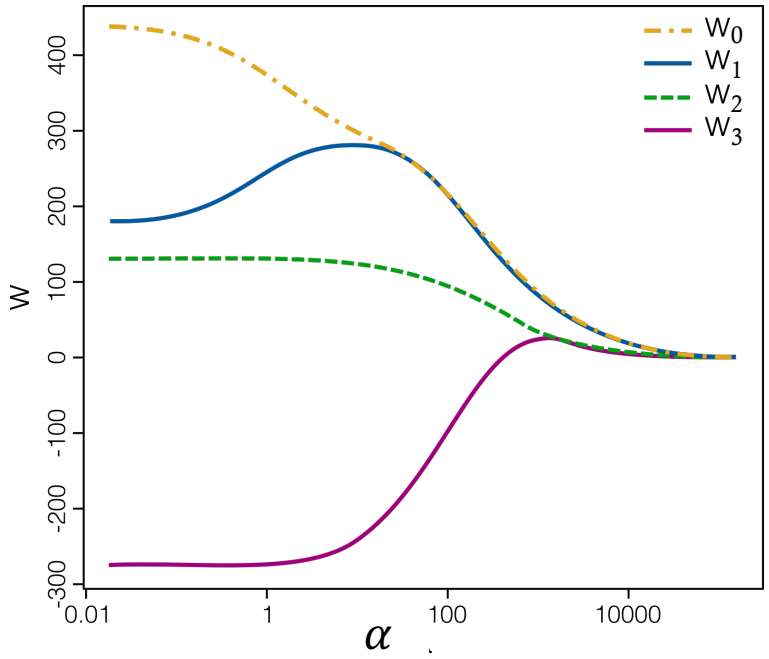
¿Para qué nos sirve?

En general, cuando la verdadera relación es lineal, la regresión lineal tiene mucha varianza. Esto principalmente ocurre cuando el **número de observaciones es cercano al número de coeficientes**. En estos casos, la regresión de Ridge funciona mejor.



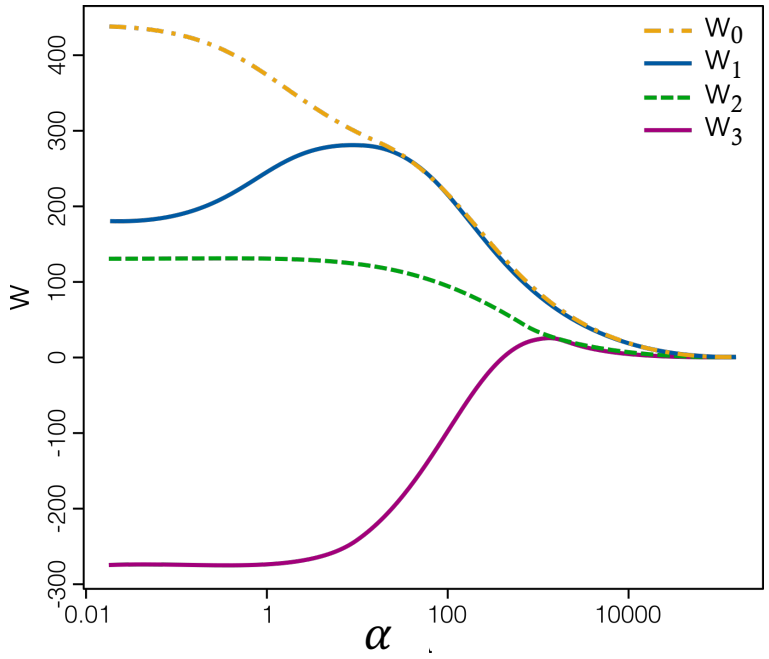
REGRESIÓN DE LASSO

La regresión de Ridge, a priori, nos parece interesante para hacer una selección de modelo, ya que, jugando con α podemos ver si algún coeficiente se hace cero:



REGRESIÓN DE LASSO

La regresión de Ridge, a priori, nos parece interesante para hacer una selección de modelo, ya que, jugando con α podemos ver si algún coeficiente se hace cero:



El problema es que los coeficientes se achican a cero, pero no se hacen cero, salvo que α sea infinito. Por lo que **no podemos eliminar atributos**.

REGRESIÓN DE LASSO

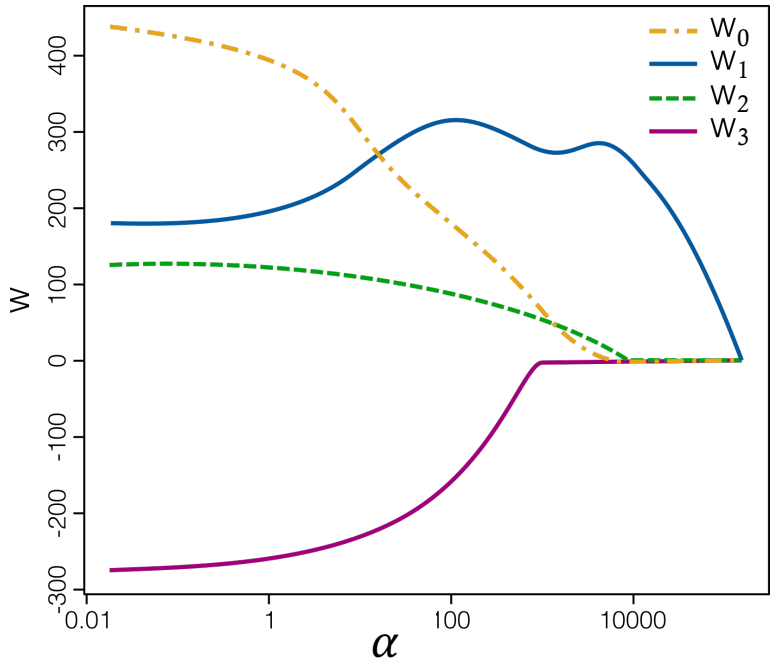
La regresión de Lasso cubre esta desventaja:

$$\sum_{i=0}^{N-1} \left(y_{[i]} - b - w_0 x_{0[i]} - \cdots - w_d x_{d[i]}\right)^2 + \alpha \sum_{j=0}^{d-1} |w_j|$$

Es decir, la regresión de Lasso usa una penalización L1, mientras que Ridge usa una penalización L2.

REGRESIÓN DE LASSO

Esta regresión cuando α crece, algunos coeficientes se hacen exactamente cero. Por lo que Lasso realiza una selección de atributos.



REGRESIÓN DE LASSO

¿Para qué nos sirve?

Para entender porque esto ocurre, debemos reescribir a las regresiones de otra forma equivalente:

Regresión Lasso

$$\underset{w}{minimizar} = \left\{ \sum_{i=0}^{N-1} \left(y_{[i]} - b - w_0 x_{0[i]} - \cdots - w_d x_{d[i]} \right)^2 \right\} \; sujeto\ a \sum_{j=0}^{d} |w_j| \leq s$$

Regresión Ridge

$$\underset{w}{minimizar} = \left\{ \sum_{i=0}^{N-1} \left(y_{[i]} - b - w_0 x_{0[i]} - \cdots - w_d x_{d[i]} \right)^2 \right\} \; sujeto\ a \sum_{j=0}^{d} w_j^2 \leq s$$

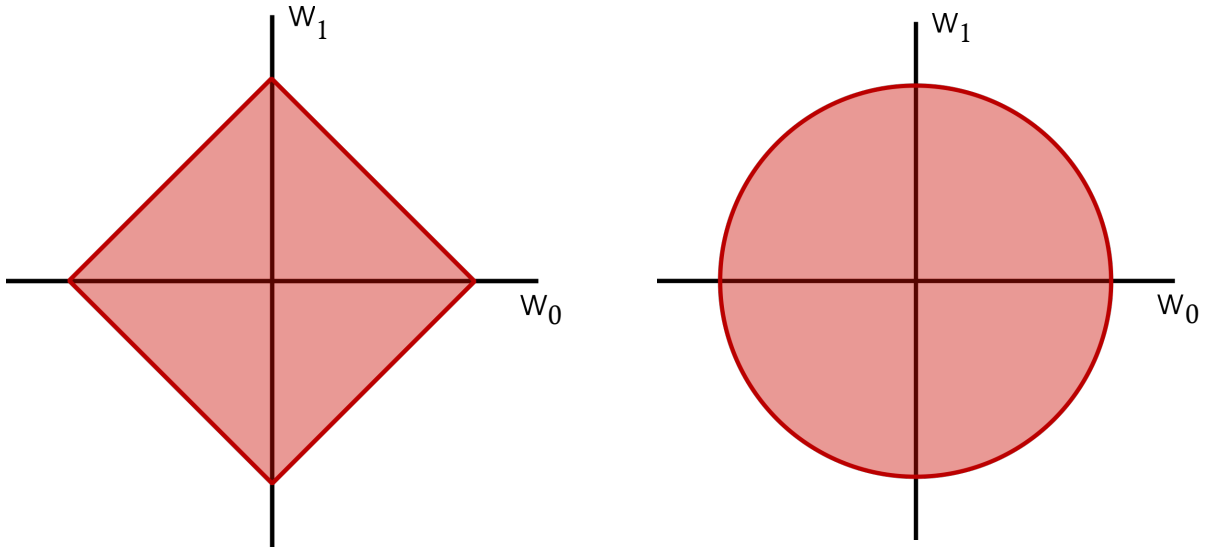
REGRESIÓN DE LASSO

¿Para qué nos sirve?

Veamos el efecto de la penalización en un caso de 2 atributos (d=2):

Regresión Lasso

$|w_0| + |w_1| \leq s$



Regresión Ridge

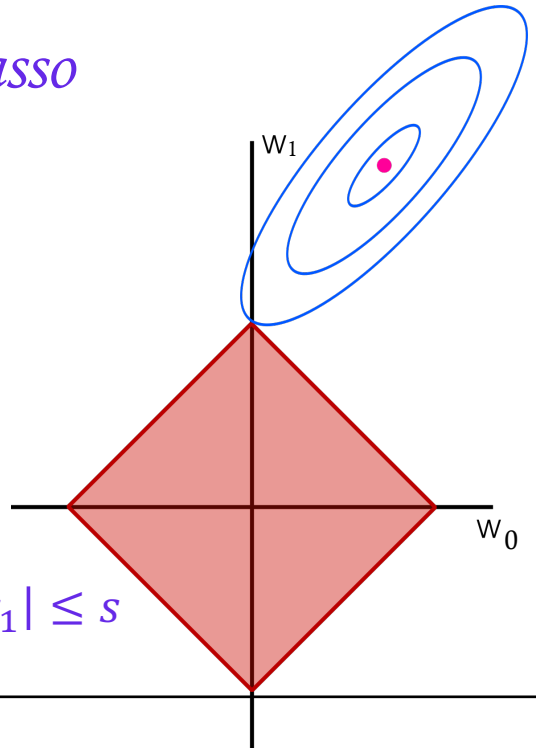
$w_0^2 + w_1^2 \leq s$

REGRESIÓN DE LASSO

¿Para qué nos sirve?

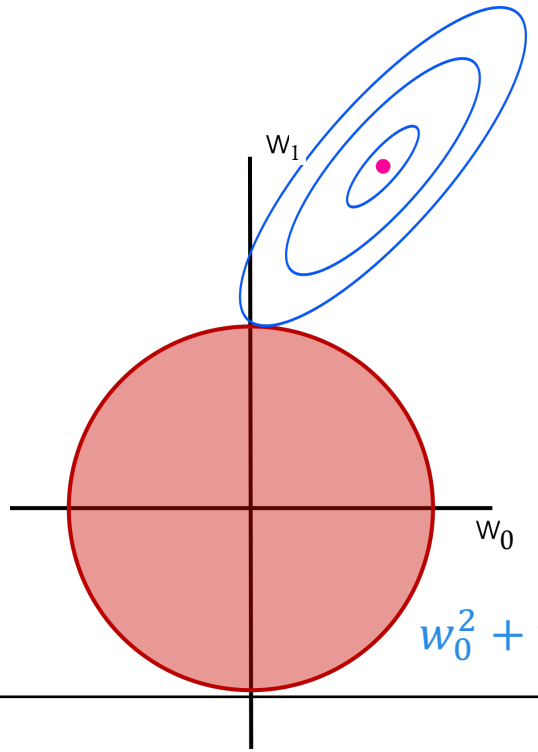
Veamos el efecto de la penalización en un caso de 2 atributos (d=2):

Regresión Lasso



$|w_0| + |w_1| \leq s$

Regresión Ridge



$w_0^2 + w_1^2 \leq s$