Day 2

2018년 12월 24일 월요일 오전 9:19

kNN

k 이웃

```
X_train, X_test, y_train, y_test = train_test_split(X, y)

model.fit(X_train, y_train)
```

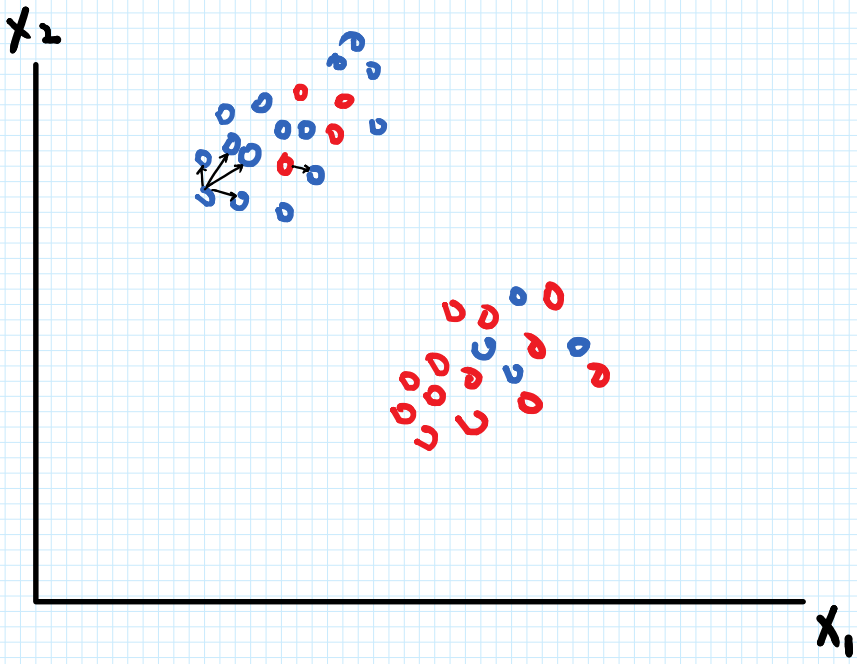
Out[27]:

거리측정 방식

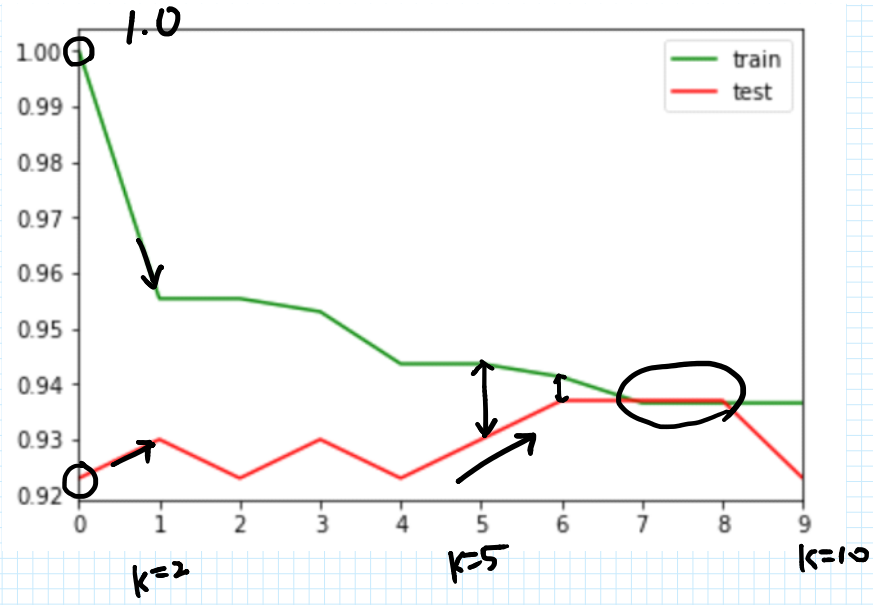
```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=1, n_neighbors=5, p=2,
                     weights='uniform')
```

이웃수

기본값



k = 1 ... 10



분류

$$y == \hat{y}$$

회귀

"결정력" 샘플별 오차 (정답 − 예측)

$$R^2 = 1 - \frac{\Sigma(y^{(i)} - \hat{y}^{(i)})^2}{\Sigma(y - \bar{y})^2}$$

$[?_0, 1.0]$

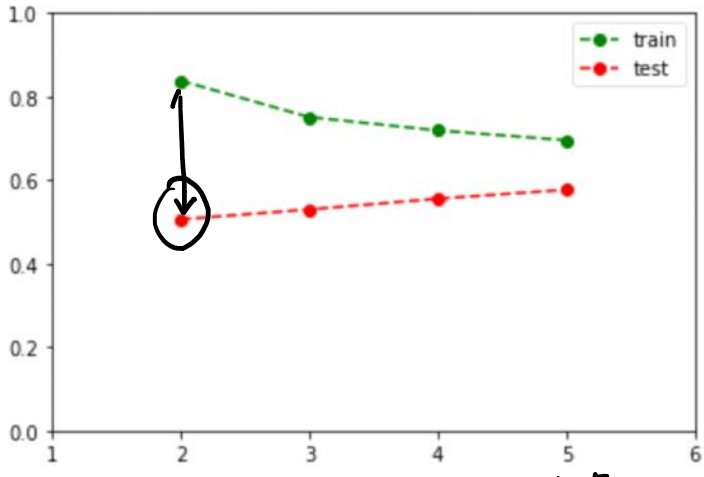
(정답, 평균)

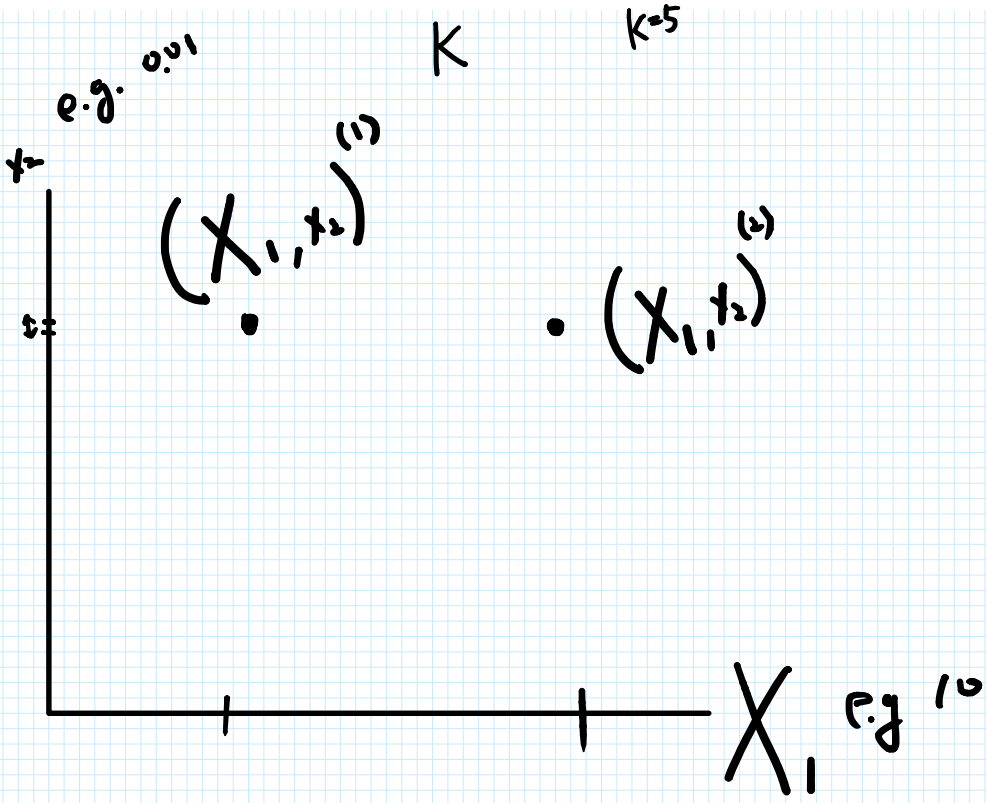
	y	ŷ	Error
1	1.0	0.9	0.1
2	1.0	1.1	-0.1
⋮			
n			

$\Sigma|Error|^2$

In [83]:
```
훈련평가.plot(ylim=(0.0, 1.0), xlim=(1, 6),
              style={'train': 'go--', 'test':'ro--'})
```

Out[83]: <matplotlib.axes._subplots.AxesSubplot at 0xe2ae860>





선형 모델

1943 MCP 뉴런

≈ "중요도"

$x_1 \times w_1$

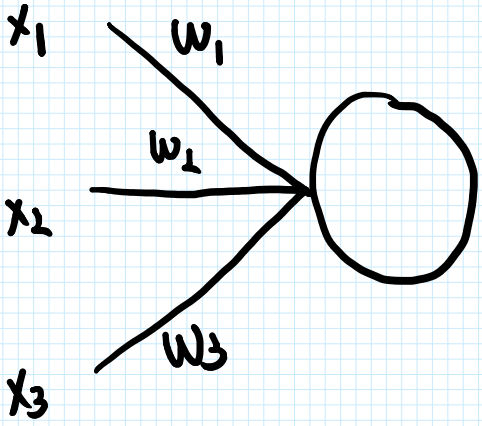
0.1 0.1 0.01

$x_1 \times w_1$ (0.1 × 0.1) → 0.01

$x_2 \times w_2$ (0.1 × 0.0) → 0.0

$x_3 \times w_3$ (0.1 × 0.0) → 0.0

0.01 → (×100) → 1.0



Σwx

$= w_1x_1 + w_2x_2 + w_3x_3$

$= z$ "정보"

e.g. KospI

$$\hat{y} = w_0x_0 + w_1x_1 + \cdots + w_px_p + b$$

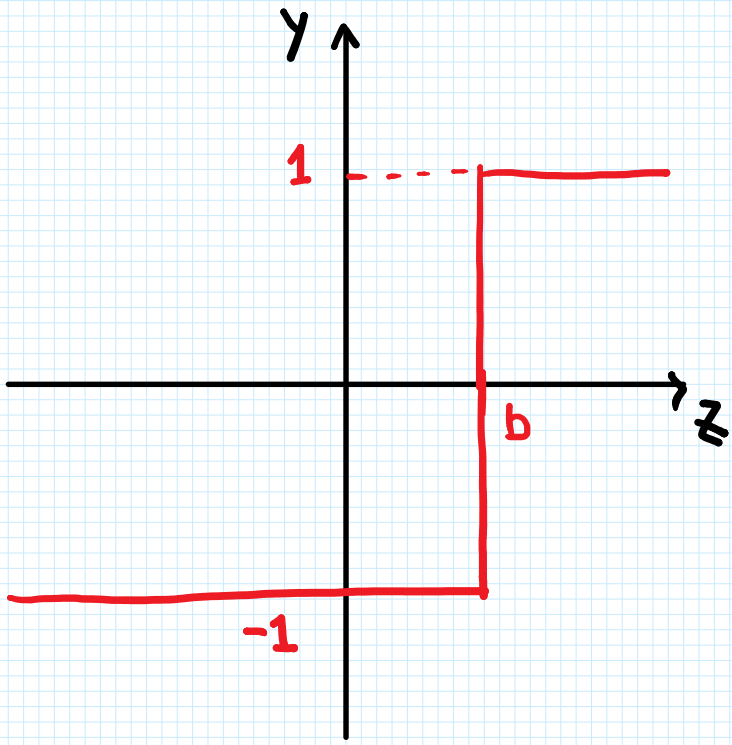
1958 퍼셉트론

$$y = \begin{cases} 1 \\ -1 \end{cases}$$

y	$\hat{y}$	$E_{rror} = y - \hat{y}$	Δw
			0
-1	1	-2	
1	-1	2	$\neq 0$
-1	1	-2	

$\begin{matrix} -1 \\ 1 \\ -1 \end{matrix}$ $\begin{matrix} 1 \\ -1 \\ 1 \end{matrix}$ $\begin{matrix} -2 \\ 2 \\ 0 \end{matrix}$

0
≠0
≠0
0



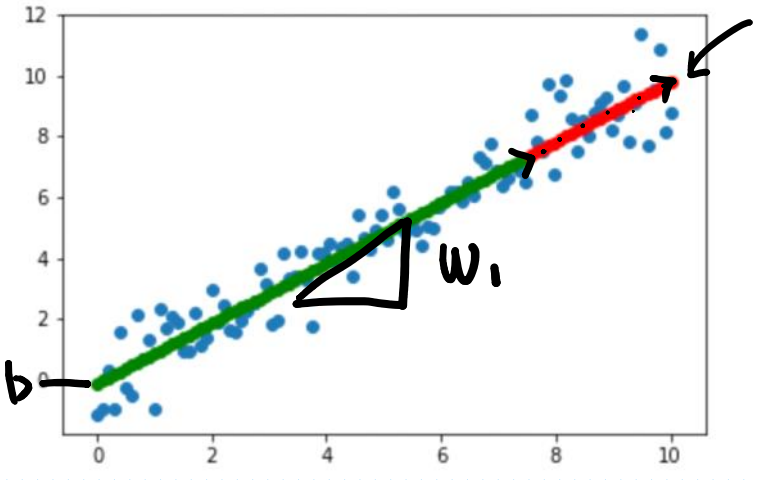
$$y = \begin{cases} 1, & z > b \\ -1, & z \le b \end{cases}$$

학습

매개변수 $[w_1 \; w_2 \; \cdots \; w_n, b]$

$x_1 \; x_2 \; \cdots \; x_n$

선형 회귀 Lin. Reg.

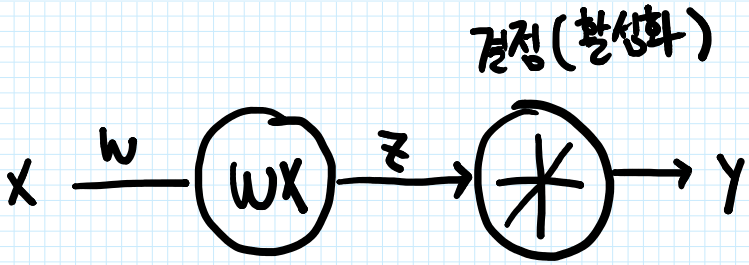


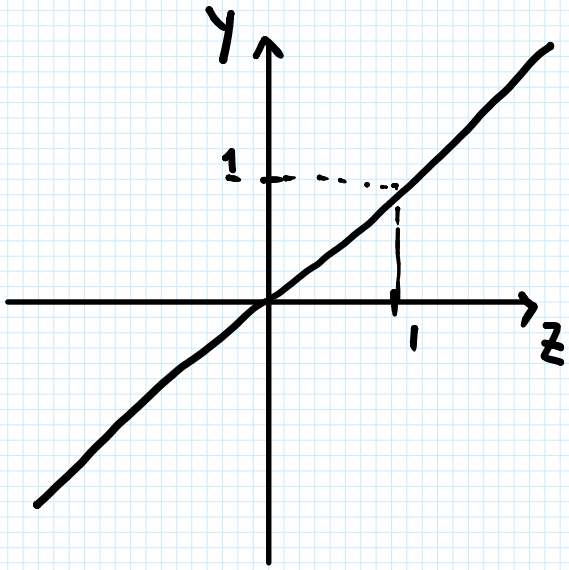
$y = x + \text{Noise}$

$y = ax + b$

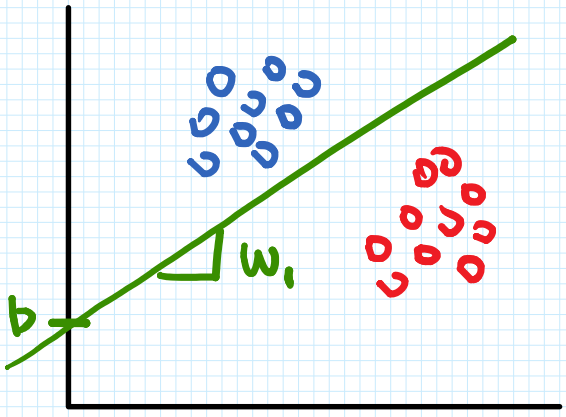
$a = 1, \; b = 0$

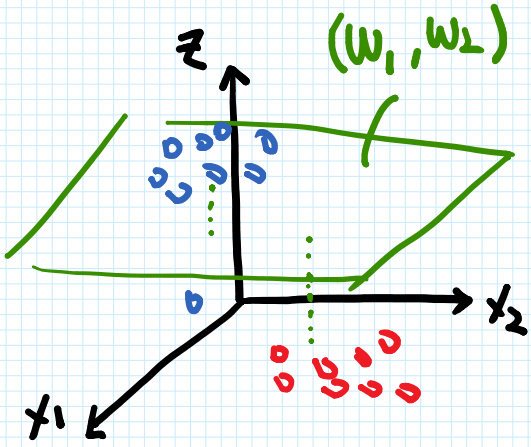
$y = w_1 x_1 + b$





$y = z$



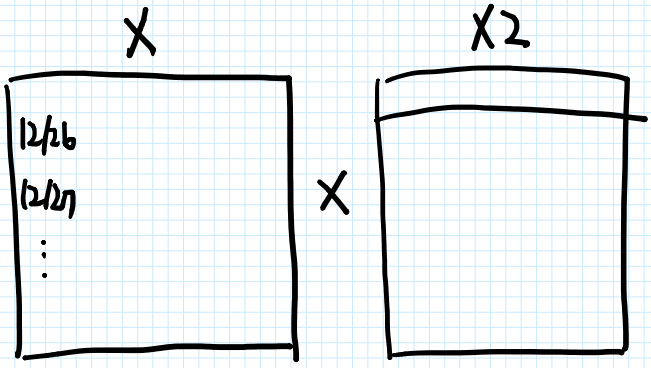


Q: 특징과 가중치의 관계 해석?

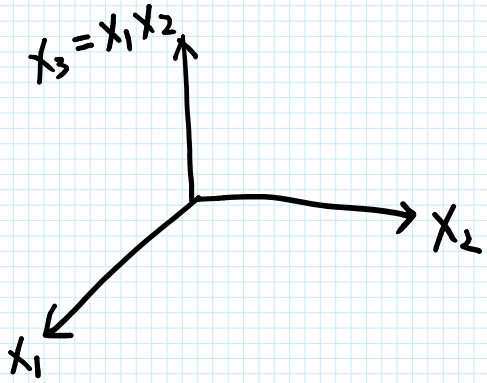
```
In [130]: pd.Series(w, index=boston.columns[1:])

Out[130]: CRIM      -0.105759
          ZN         0.054570
          INDUS      0.041421
          CHAS       3.221044
          NOX      -18.033190
          RM         3.702777
          AGE        0.004706
          DIS       -1.367117
          RAD        0.307633
          TAX       -0.011551
          PTRATIO   -1.005021
          B          0.011575
          LSTAT     -0.497286
          dtype: float32
```

선형모델 성능과 데이터 특징

1)



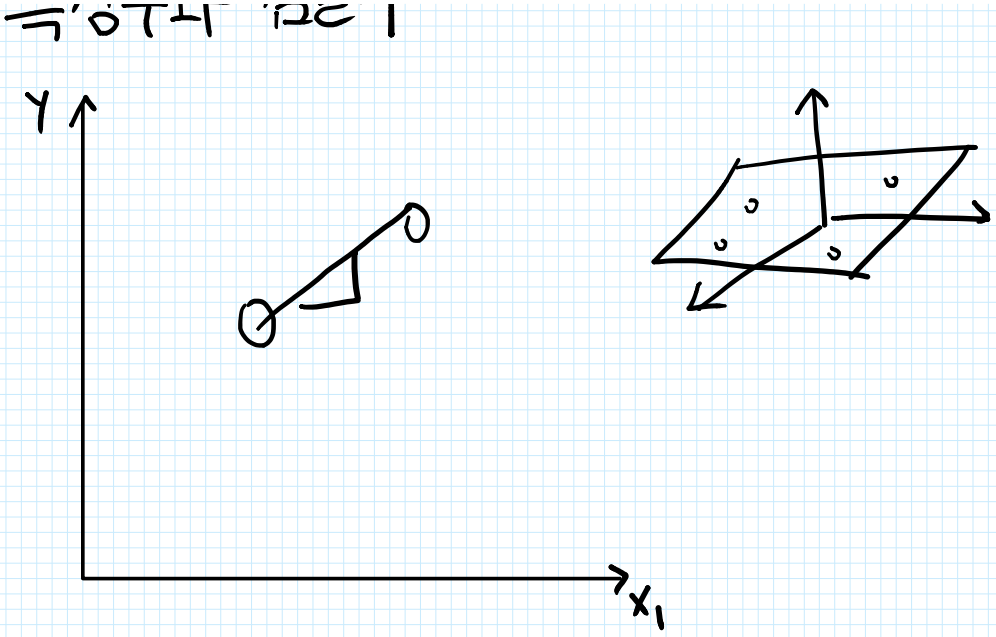
2) 데이터 공학



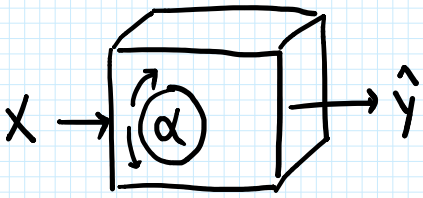
$$\overbrace{x_1\ x_2}^{2} \rightarrow \overbrace{x_1^2\ \ x_1x_2\ \ x_2^2}^{n}$$

$w_1\ w_2\ w_3\ w_4\ w_5$

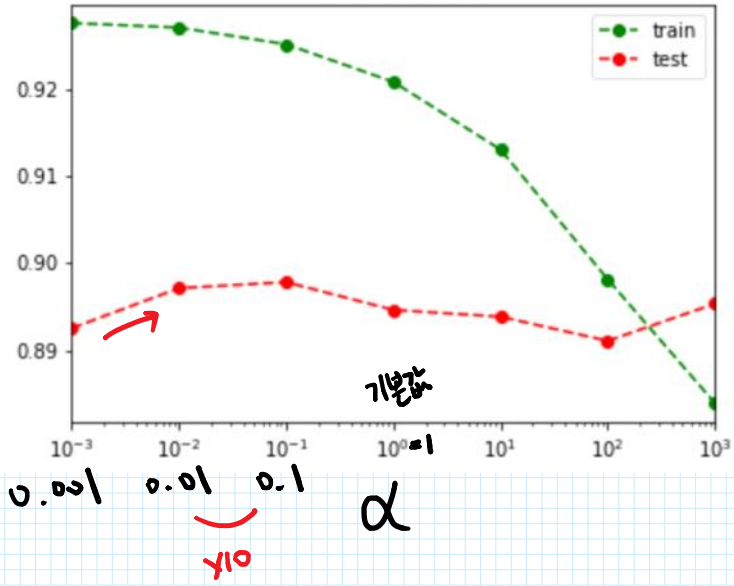
특징수와 샘플수



Ridge




```
In [151]: 훈련평가.plot(logx=True, style=train_test_styles)
Out[151]: <matplotlib.axes._subplots.AxesSubplot at 0x106d16a0>
```

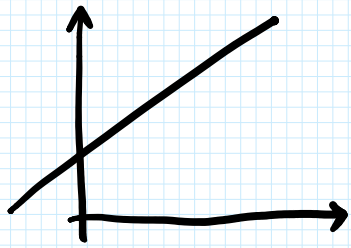


선형

모델 표현력과 가중치의 크기

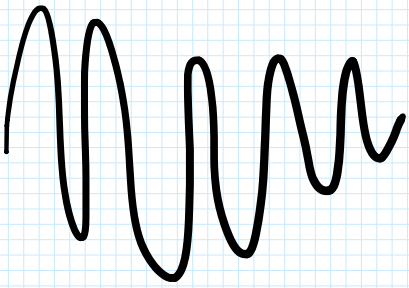
$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

$$y = w_1 x_1 + b, \quad w_2 = \cdots = w_n = 0$$



⇒ 과소

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$



⇒ 과대

$$w \rightarrow w \neq 0$$

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

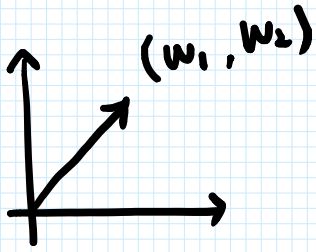
"적합"

fit

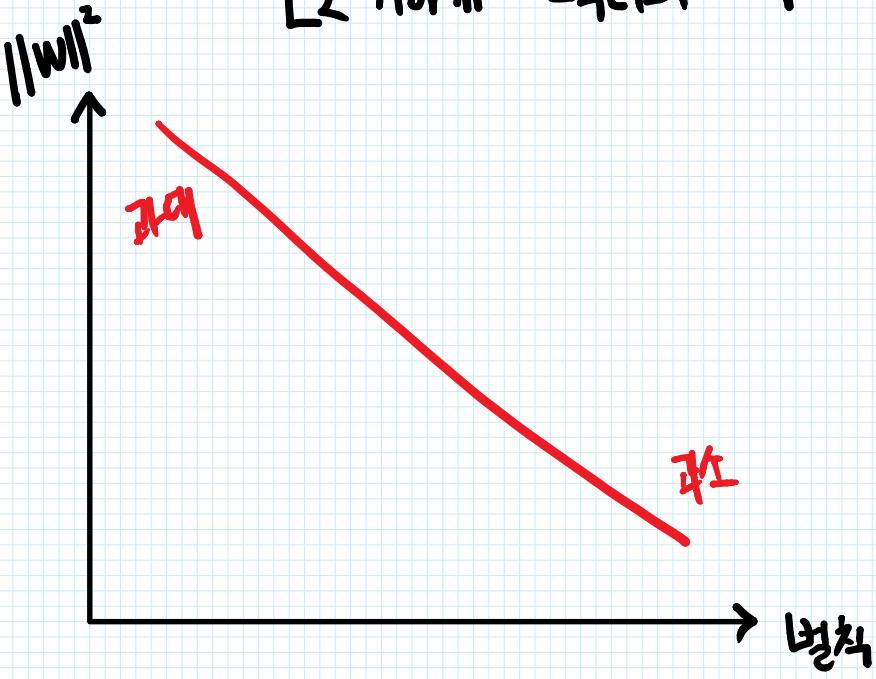
벌칙 $w_1^2 + w_2^2 + \cdots + w_n^2$

$= \sum_i w_i^2$

$= \|w\|^2$



L2 norm 벡터의 크기



Ridge ⇐ Lin. Reg. + 벌칙

$\alpha\|w\|^2$

가중치 크기

e.g. $\|w\|^2 = 100$

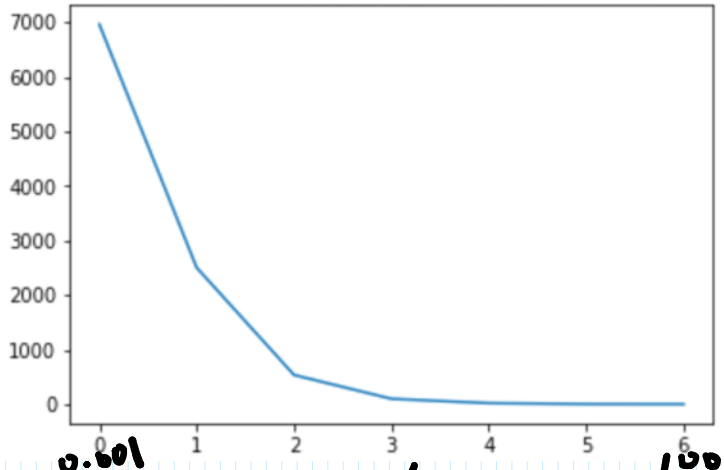
$\alpha = 0.01 \rightarrow \alpha\|w\|^2 = 1$ 자유

$\alpha = 1.0 \rightarrow \alpha\|w\|^2 = 100$

$\alpha = 100 \rightarrow \alpha\|w\|^2 = 10K$ 제한

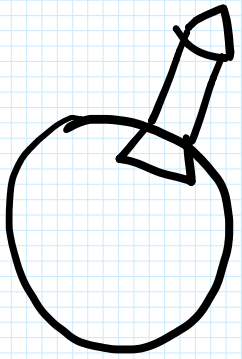
In [161]:
```
가중치배열 = np.array(가중치목록)
가중치크기 = np.sum(가중치배열**2, axis=1)
plt.plot(가중치크기)
```

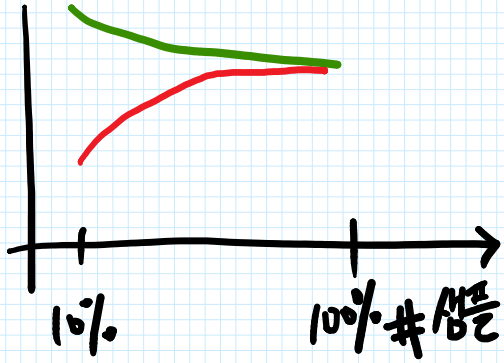
Out[161]: [<matplotlib.lines.Line2D at 0x130f6898>]



0.001 α 1000

	α	w_0	w_1	...	w_n	$\|w\|^2$
1	0.001					
2	0.01					
3	0.1					
4	1.0					
5	10					
6	100					
7	100					




1%
10%