Chap 4. 모델 훈련 (Model Training)

Seolyoung Jeong, Ph.D.

경북대학교 IT 대학

Contents

4.1 선형 회귀
- 4.1.1 정규방정식
- 4.1.2 계산 복잡도

4.2 경사 하강법
- 4.2.1 배치 경사 하강법
- 4.2.2 확률적 경사 하강법
- 4.2.3 미니배치 경사 하강법

4.3 다항 회귀

4.4 학습 곡선

4.5 규제가 있는 선형 모델
- 4.5.1 릿지 회귀
- 4.5.2 라쏘 회귀
- 4.5.3 엘라스틱넷
- 4.5.4 조기 종료

4.6 로지스틱 회귀
- 4.6.1 확률 추정
- 4.6.2 훈련과 비용 함수
- 4.6.3 결정 경계
- 4.6.4 소프트맥스 회귀

모델 훈련

- 머신러닝 모델, 훈련 알고리즘 → 블랙박스 취급
- 실제로 어떻게 작동하는지는 모름

- 어떻게 작동하는지 잘 이해하고 있으면...
- 적절한 모델, 올바른 훈련 알고리즘, 작업에 맞는 좋은 하이퍼파라미터를 빠르게 찾을 수 있음
- 디버깅이나 에러를 효율적으로 분석 가능
- → 특히 신경망을 이해, 구축, 훈련시키는데 필수

- **모델을 훈련시킨다. =**
  모델이 훈련세트에 가장 잘 맞도록 모델 파라미터를 설정한다.
  - 두 가지 방법) 직접 계산 가능한 공식 사용 /
    반복적 최적화 방식(경사 하강법)을 사용해서 모델 파라미터를 조금씩 바꾸면서 비용 함수를 훈련 세트에 대해 최소화

- **먼저, 모델이 훈련 데이터에 얼마나 잘 들어맞는지 측정**
- **성능 측정 지표 : 평균 제곱근 오차 (RMSE)**
- **즉, RMSE를 최소화하는 θ를 찾아야 함**
  - 실제로는 RMSE보다 평균제곱오차(MSE)를 최소화하는 것이 같은 결과를 내면서 더 간단
  - 선형 회귀 모델의 MSE 비용 함수

$$\mathrm{MSE}(\mathbf{X}, h_\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{\theta}^{\boldsymbol{T}} \cdot \boldsymbol{X}^{(i)} - \boldsymbol{y}^{(i)}\right)^2$$

4.1 선형 회귀 (Linear Regression)

- 가장 간단한 모델 중 하나
- '삶의 만족도' 선형 회귀 모델

$$\text{삶의만족도} = \theta_0 + \theta_1 \text{X1인당 GDP}$$

- 선형 모델
  - 예측값 = 입력 특성의 가중치 합 + 편향(또는 절편)이라는 상수

$$\hat{y} = \boldsymbol{\theta_0} + \boldsymbol{\theta_1 x_1} + \boldsymbol{\theta_2 x_2} + \cdots + \boldsymbol{\theta_n x_n}$$

$\hat{y}$: 예측값, n: 특성 수, x_i: i번째 특성값($x_0 = 1$),
θ_j: j번째 모델 파라미터(편향(θ_0)과 가중치($\theta_1, \theta_2 \ldots \theta_n$) 포함)

- 벡터 형태로 표현

$$\hat{y} = \boldsymbol{h_\theta(x) = \theta^T \cdot X}$$

→ 선형 회귀 모델의 예측

4.1.1 정규방정식

- 비용함수를 최소화하는 θ 값을 찾기 위한 해석적인 방법 (수학공식) : 정규방정식

$$\hat{\theta} = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot y$$

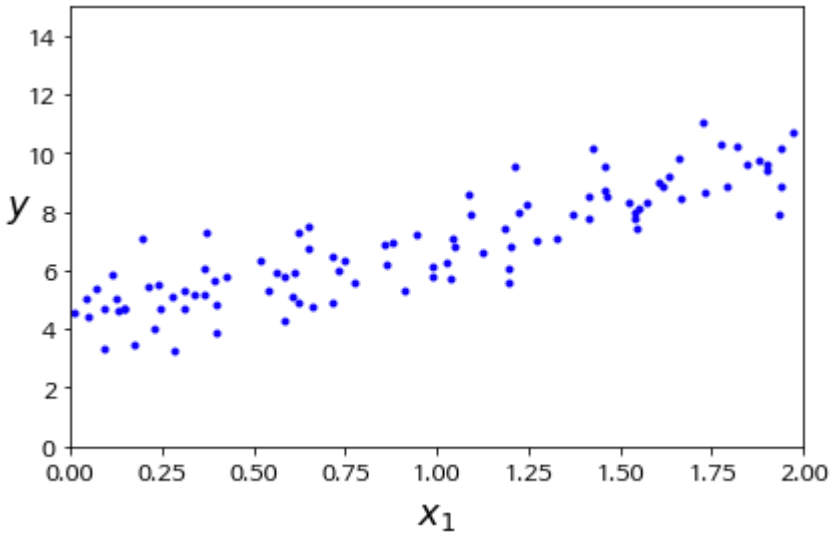
$\hat{\theta}$: 비용함수를 최소화시키는 θ값 벡터

y : $y^{(1)}$부터 $y^{(m)}$까지 포함하는 타깃 벡터

- 테스트 위해 임의의 선형 데이터 생성

```
import numpy as np

X = 2 * np.random.rand(100, 1)
y = 4 + 3 * X + np.random.randn(100, 1)
```

```
plt.plot(X, y, "b.")
plt.xlabel("$x_1$", fontsize=18)
plt.ylabel("$y$", rotation=0, fontsize=18)
plt.axis([0, 2, 0, 15])
plt.show()
```



정규방정식을 사용해 $\hat{\theta}$ 계산

- 역행렬 계산 : numpy 선형대수 모듈 (np.linalg)의 inv() 함수
- 행렬 곱셈 : dot() 함수 사용

```
X_b = np.c_[np.ones((100, 1)), X]  # 모든 샘플에 x0 = 1을 추가합니다.
theta_best = np.linalg.inv(X_b.T.dot(X_b)).dot(X_b.T).dot(y)
```

- 값 확인 → 기대값 : θ_0=4, θ_1=3 (노이즈 때문에 정확하지 않음)

```
theta_best
```

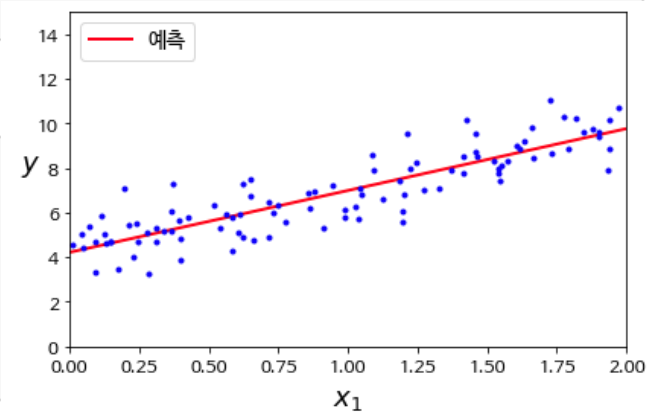
```
array([[4.21509616],
       [2.77011339]])
```

- $\hat{\theta}$ 을 사용해 예측

```
X_new = np.array([[0], [2]])
X_new_b = np.c_[np.ones((2, 1)), X_new]  # 모든 샘플에 x0 = 1을 추가합니다.
y_predict = X_new_b.dot(theta_best)
y_predict
```

```
array([[4.21509616],
       [9.75532293]])
```

```
plt.plot(X_new, y_predict, "r-", linewidth=2, label="예측")
plt.plot(X, y, "b.")
plt.xlabel("$x_1$", fontsize=18)
plt.ylabel("$y$", rotation=0, fontsize=18)
plt.legend(loc="upper left", fontsize=14)
plt.axis([0, 2, 0, 15])
plt.show()
```



◆ 사이킷런의 LinearRegression 사용하는 방법

```
from sklearn.linear_model import LinearRegression
lin_reg = LinearRegression()
lin_reg.fit(X, y)
lin_reg.intercept_, lin_reg.coef_
```

```
(array([4.21509616]), array([[2.77011339]]))
```

```
lin_reg.predict(X_new)
```

```
array([[4.21509616],
       [9.75532293]])
```

4.1.2 계산 복잡도

- 정규방정식은 $(n+1)$x$(n+1)$ 크기가 되는 $X^T \cdot X$의 역행렬 계산 (n은 특성 수)

- 역행렬 계산 복잡도 : $O(n^{2.4}) \sim O(n^3)$
- → 특성 수가 2배로 늘어나면 계산 시간이 대략 5.3~8배로 증가

- 훈련 세트의 샘플 수에는 선형적으로 증가 $O(m)$
- 메모리 공간이 허락된다면 큰 훈련 세트도 효율적으로 처리 가능

- 선형 회귀 모델 예측 빠름.
- 예측 계산 복잡도는 샘플수와 특성수에 선형적
- → 예측하려는 샘플이 두배 증가, 걸리는 시간 두배 증가

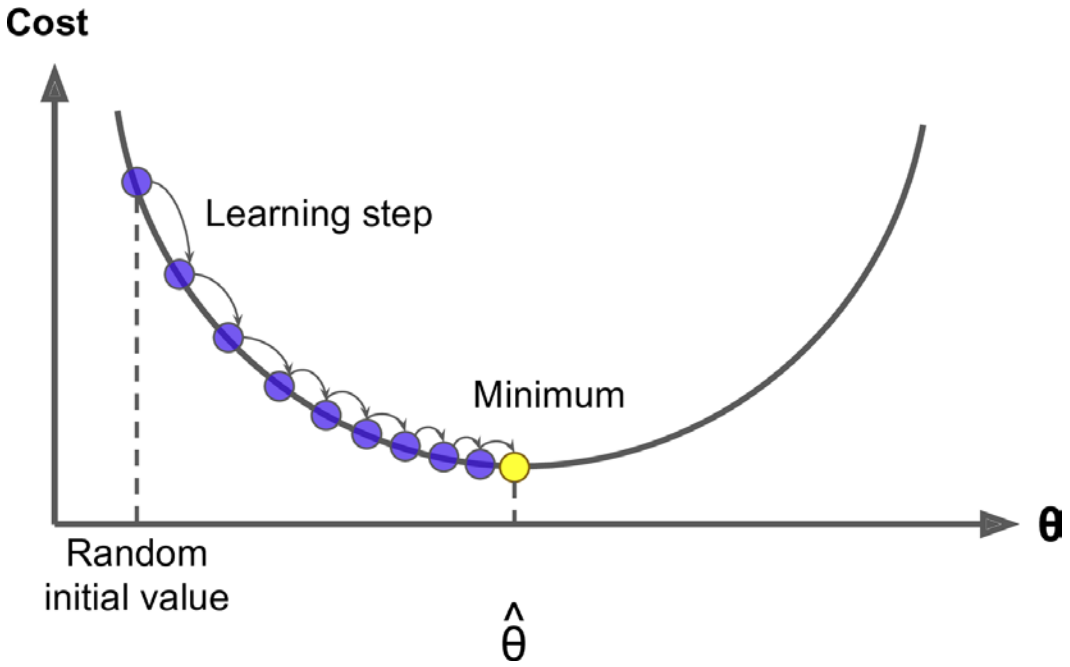
4.2 경사 하강법

- **또 다른 방법으로 선형 회귀 모델 훈련**

- **특성이 매우 많고 훈련 샘플이 너무 많아 메모리에 모두 담을 수 없을 때 적합**

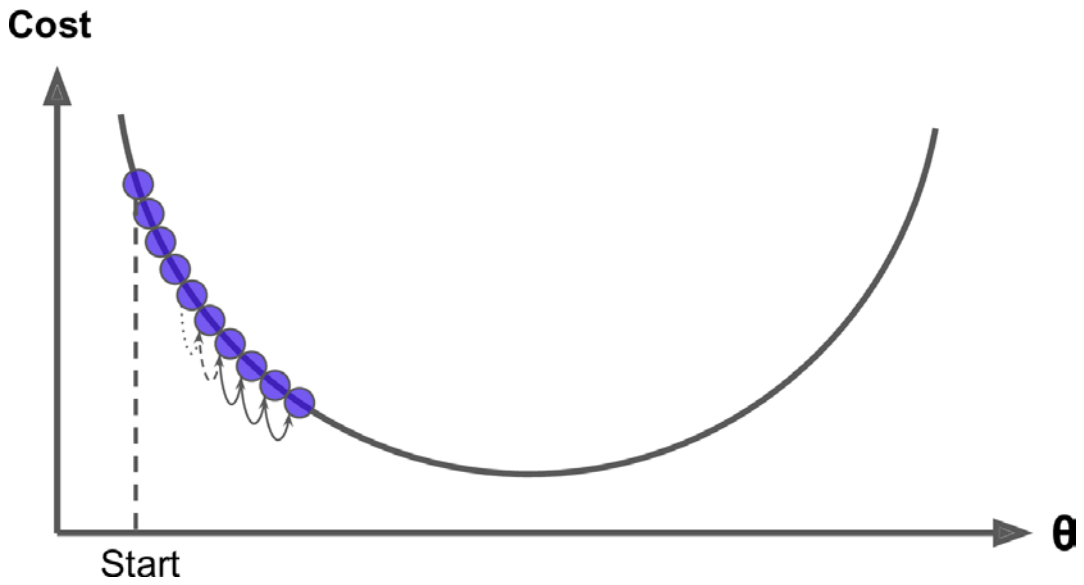
- **경사하강법(Gradient Descent)**
  - 여러 종류의 문제에서 최적의 해법을 찾을 수 있는 매우 일반적인 최적화 알고리즘
  - 기본 아이디어) 비용 함수를 최소화하기 위해 반복해서 파라미터 조정
  - 파라미터 벡터 θ 에 대해 비용 함수의 gradient를 계산하고, gradient가 감소하는 방향으로 반복적으로 θ를 수정
  - gradient=0 이면 최소값에 도달 완료

경사 하강법

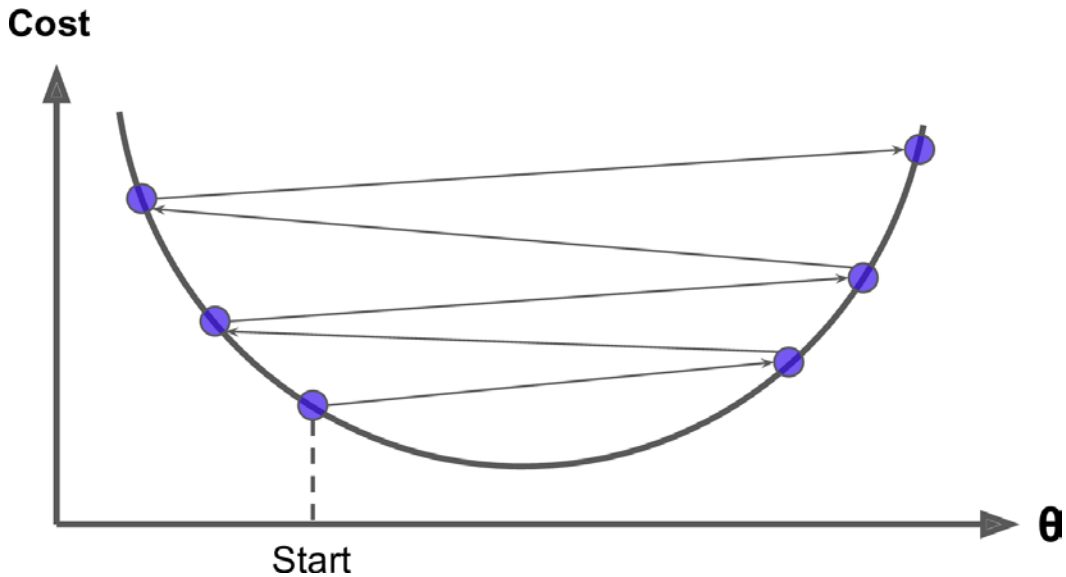


- Learning step의 크기는 학습률 하이퍼파라미터로 결정됨
- 학습률이 너무 작으면 알고리즘이 수렴하기 위해 반복을 많이 진행. 시간이 오래 걸림
- 학습률이 너무 크면 골짜기를 가로질러 반대편으로 건너뛰게 되어 이전보다 더 높은 곳으로 올라갈지도...

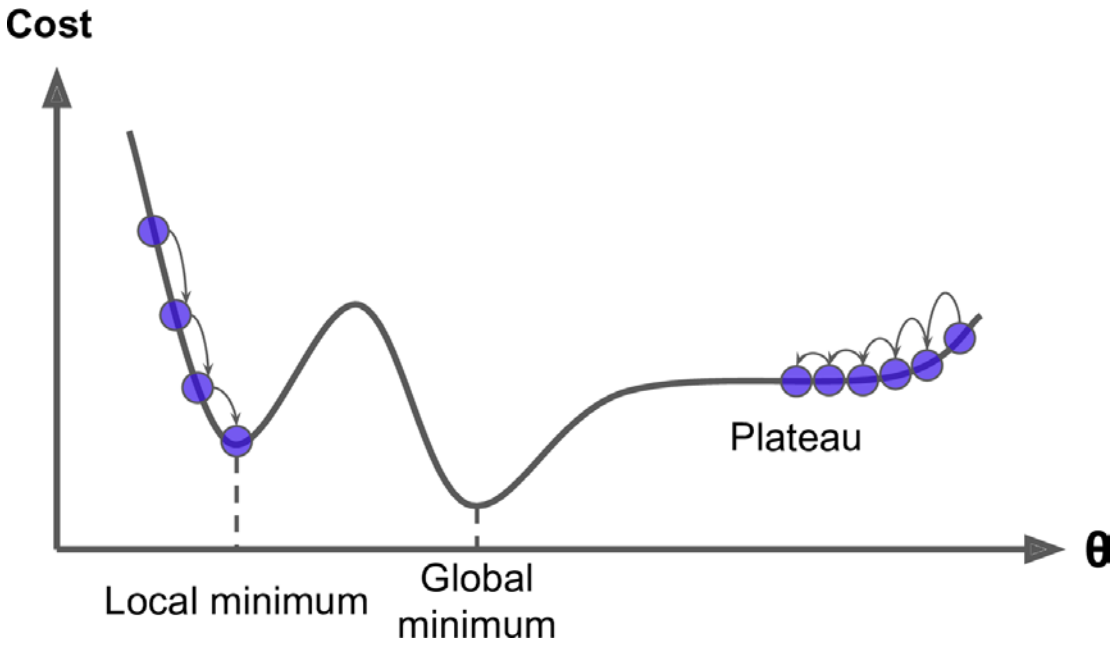
- 학습률이 너무 작을 때



- 학습률이 너무 클 때

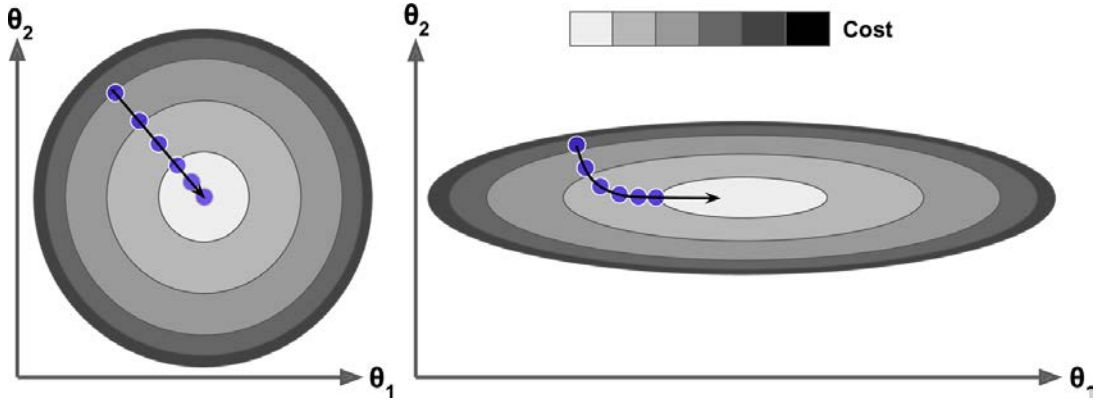


- **모든 비용 함수가 매끈한 그릇 같지는 않음**
  - 패인 곳, 산마루, 평지 등 특이한 지형
  - 최솟값으로 수렴하기 매우 어려움
- **경사 하강법의 문제점 예)**
  - 왼쪽에서 시작 → 전역 최솟값보다 덜 좋은 지역 최솟값에 수렴
  - 오른쪽에서 시작 → 평탄한 지역을 지나기 위해 시간이 오래 걸리고, 일찍 멈추게 됨



- **선형 회귀를 위한 MSE 비용 함수**
  - 곡선에서 어떤 두 점을 선택해 선을 그어도 곡선을 가로지르지 않는 볼록 함수 (convex function)
  - 지역 최솟값이 없고, 하나의 전역 최솟값만 있음
  - 연속된 함수. 기울기가 갑자기 변하지 않음

  → 경사 하강법이 전역 최솟값에 가깝게 접근할 수 있음을 보장
  (학습률이 너무 높지 않고, 충분한 시간이 주어진다면...)
- **특성의 스케일이 매우 다르면 길쭉한 모양**
  - (왼쪽) 특성1=특성2 스케일 : 최솟값으로 곧장 진행. 빠르게 도달
    (오른쪽) 특성1<특성2 스케일 : 시간 오래 걸림



  - scaling 필요 : StandardScaler() 함수 사용

- **모델 훈련이란...**
  - (훈련 세트에서) 비용 함수를 최소화하는 모델 파라미터의 조합을 찾음
  - 모델이 가진 파라미터가 많을수록 공간의 차원을 커지고 검색 어려워짐

4.2.1 배치 경사 하강법

- **경사 하강법 구현**
  - 각 모델 파라미터 θ_j에 대해 비용 함수의 gradient 계산
  - θ_j가 조금 변경될 때 비용 함수가 얼마나 바뀌는지 계산
  - 편도함수(partial derivative)

- **파라미터 θ_j에 대한 비용 함수의 편도 함수**
  - 모든 차원에 대해 기울기 확인

$$\frac{\partial}{\partial \theta_j} \mathrm{MSE}(\mathbf{X}, \mathrm{h}_\theta) = \frac{2}{M} \sum_{i=1}^{m} \left(\theta^T \cdot \boldsymbol{X}^{(i)} - y^{(i)} \right) x_j^{(i)}$$

  - 각각 계산하는 대신 한꺼번에 계산
  - 그래디언트 벡터 : 비용 함수의 편도함수를 모두 담음

$$\nabla_\theta \mathrm{MSE}(\theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_0} \mathrm{MSE}(\theta) \\ \frac{\partial}{\partial \theta_1} \mathrm{MSE}(\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_n} \mathrm{MSE}(\theta) \end{pmatrix} = \frac{2}{m} \mathbf{X}^T \cdot (\mathbf{X} \cdot \theta - \mathbf{y})$$

매 경사 하강법 스텝에서 전체 훈련 세트 X에 대해 계산
→ 배치 경사 하강법
매우 큰 훈련세트에서 아주 느림
하지만, 특성 수에 민감하지 않음

다음에 내려가는 스텝 크기 결정

- θ 에서 $\nabla_\theta MSE$를 뺌
- 학습률 η 사용 → 이전의 그래디언트 벡터 * η
- 경사하강법의 스텝

$$\theta^{(next\ step)} = \theta - \eta \nabla_\theta MSE(\theta)$$

- 알고리즘 구현

```
eta = 0.1
n_iterations = 1000
m = 100
theta = np.random.randn(2,1)

for iteration in range(n_iterations):
    gradients = 2/m * X_b.T.dot(X_b.dot(theta) - y)
    theta = theta - eta * gradients
```

```
theta
```

```
array([[4.21509616],
       [2.77011339]])
```

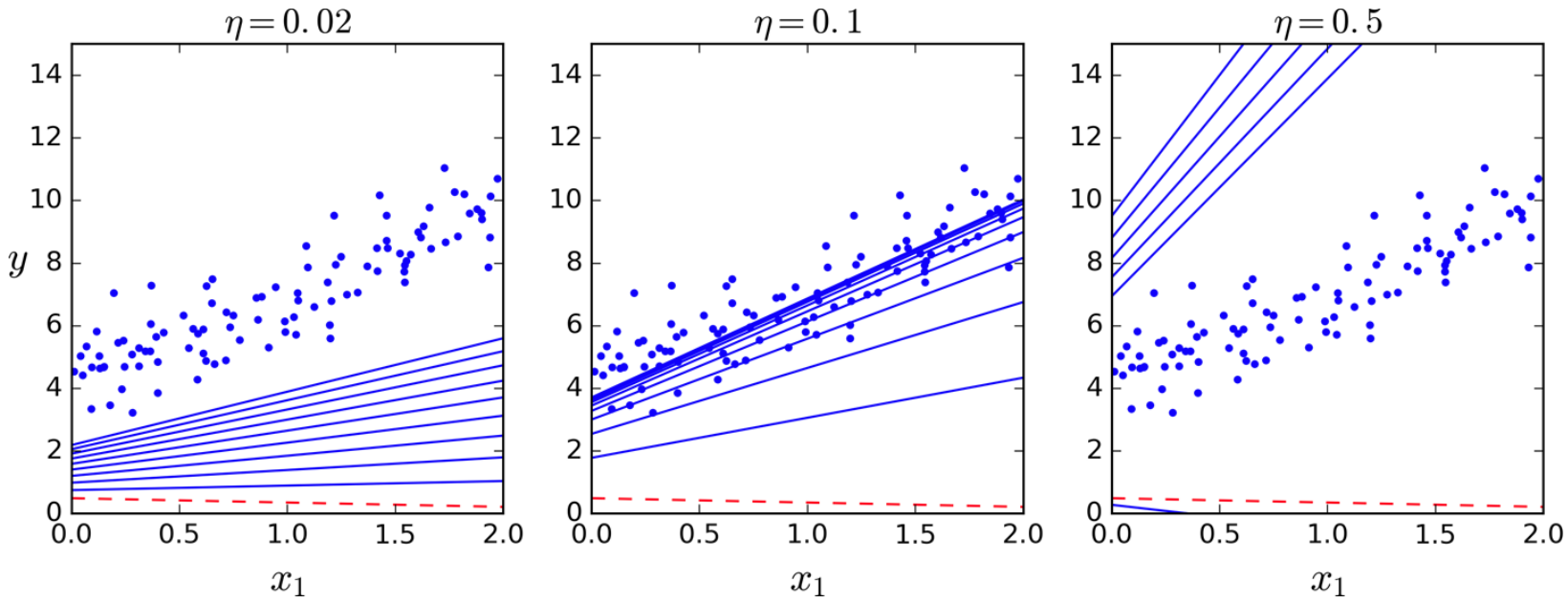
정규방정식으로 찾은 것과 동일

```
X_new_b.dot(theta)
```

```
array([[4.21509616],
       [9.75532293]])
```

◆ 학습률 η 변경

- 세 가지 다른 학습률 사용하여 진행한 경사 하강법 스텝 처음 10개 (점선은 시작점)



학습률이 너무 낮음
시간 오래 걸림

학습률 적당
반복 몇 번 만에 최적점 수렴

학습률이 너무 높음
알고리즘이 이리저리 널뛰면서
스텝마다 최적점에서 점점 더
멀어져 발산

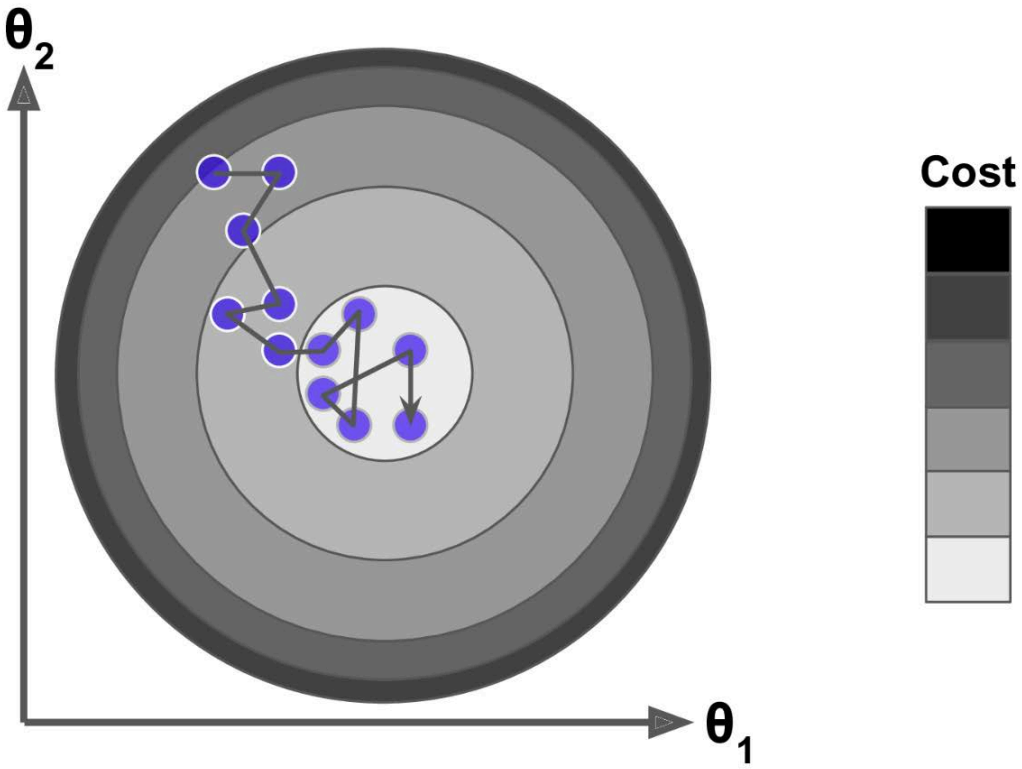
BGD 구현

```
theta_path_bgd = []

def plot_gradient_descent(theta, eta, theta_path=None):
    m = len(X_b)
    plt.plot(X, y, "b.")
    n_iterations = 1000
    for iteration in range(n_iterations):
        if iteration < 10:
            y_predict = X_new_b.dot(theta)
            style = "b-" if iteration > 0 else "r--"
            plt.plot(X_new, y_predict, style)
        gradients = 2/m * X_b.T.dot(X_b.dot(theta) - y)
        theta = theta - eta * gradients
        if theta_path is not None:
            theta_path.append(theta)
    plt.xlabel("$x_1$", fontsize=18)
    plt.axis([0, 2, 0, 15])
    plt.title(r"$\eta = {}$".format(eta), fontsize=16)
```

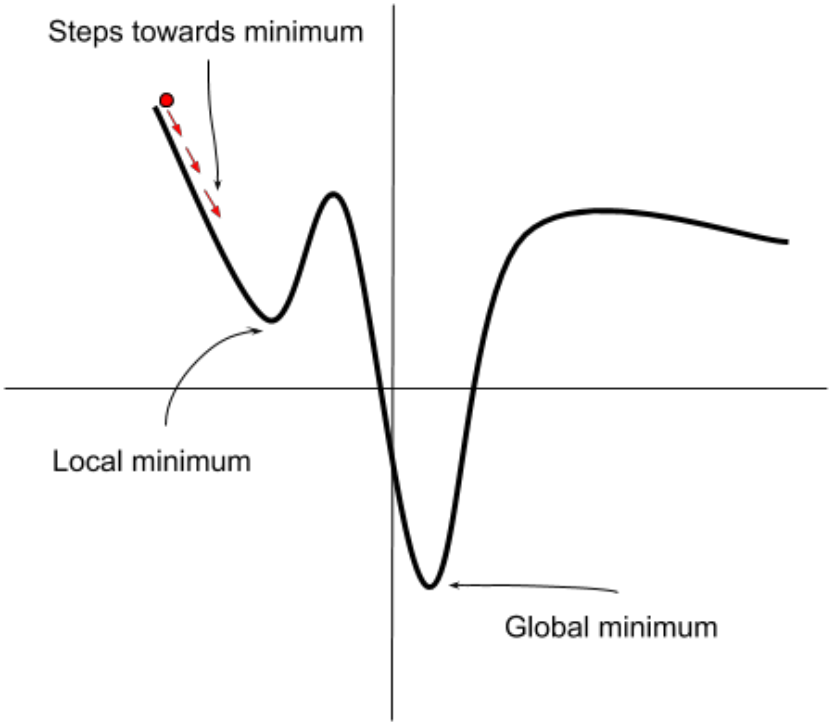
```
np.random.seed(42)
theta = np.random.randn(2,1)  # random initialization

plt.figure(figsize=(10,4))
plt.subplot(131); plot_gradient_descent(theta, eta=0.02)
plt.ylabel("$y$", rotation=0, fontsize=18)
plt.subplot(132); plot_gradient_descent(theta, eta=0.1, theta_path=theta_path_bgd)
plt.subplot(133); plot_gradient_descent(theta, eta=0.5)

plt.show()
```

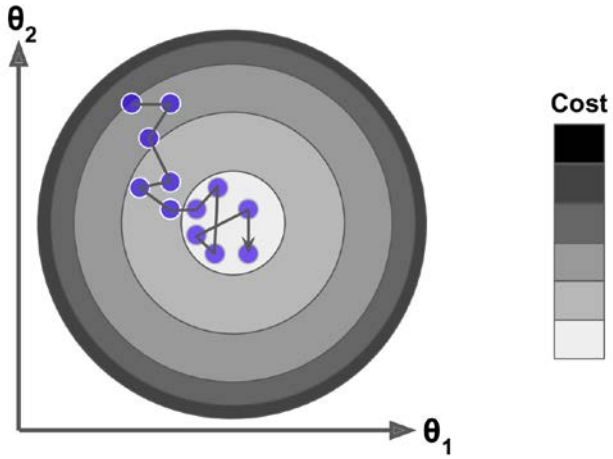
4.2.2 확률적 경사 하강법 (SGD)

- SGD(Stochastic gradient descent)는 매 step에서 한 개의 샘플을 무작위로 선택하고 하나의 샘플에 대한 gradient를 계산
- 매 반복에서 하나의 샘플만 있으면 되므로 매우 큰 훈련세트도 빠르게 훈련시킬 수 있음
- 단점) 확률적이기 때문에 batch gradient descent 보다 불안정



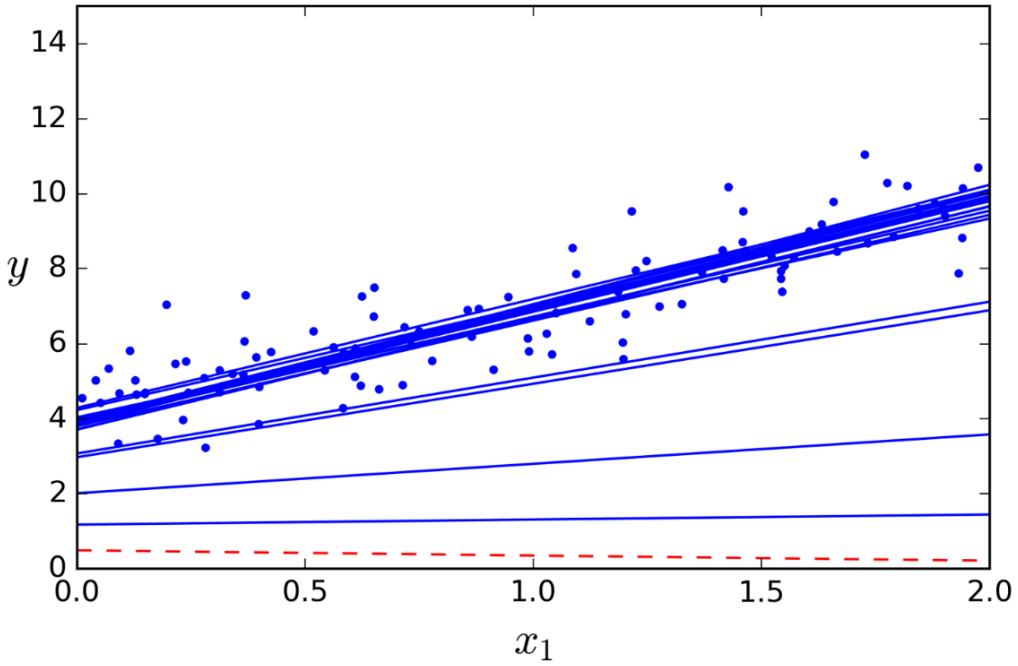
- 만일 cost function이 오른쪽 그림과 같이 매우 불규칙하다면 알고리즘이 지역 최소값을 건너뛸 수 있도록 도와주므로 SGD가 BGD보다 전역 최솟값을 찾을 가능성이 더 높음
- 무작위성은 지역 최소값을 건너뛸 수 있어 좋지만 전역 최소값에 정확히 다다르지 못한다는 단점





- 딜레마를 해결하기 위해 학습률을 점진적으로 감소시키는 방법을 사용
- 매 반복에서 학습률을 결정하는 함수 : **learning (rate) schedule**

- 학습률이 너무 빨리 줄어들면 지역 최솟값에 빠지거나 최솟값까지 가는 도중 멈출 수 있음
- 반면 너무 천천히 줄어들면 최솟값 주변을 오랫동안 머물거나 훈련을 일찍 중지시켜 지역 최솟값에 머무르게 할 수 있음



SGD 구현

```
theta_path_sgd = []
m = len(X_b)
np.random.seed(42)
```

```
n_epochs = 50
t0, t1 = 5, 50  # 학습 스케줄 하이퍼파라미터 learning schedule hyperparameters

def learning_schedule(t):
    return t0 / (t + t1)

theta = np.random.randn(2,1)  # 무작위 초기화

for epoch in range(n_epochs):
    for i in range(m):
        if epoch == 0 and i < 20:
            y_predict = X_new_b.dot(theta)
            style = "b-" if i > 0 else "r--"
            plt.plot(X_new, y_predict, style)
        random_index = np.random.randint(m)
        xi = X_b[random_index:random_index+1]
        yi = y[random_index:random_index+1]
        gradients = 2 * xi.T.dot(xi.dot(theta) - yi)
        eta = learning_schedule(epoch * m + i)
        theta = theta - eta * gradients
        theta_path_sgd.append(theta)

plt.plot(X, y, "b.")
plt.xlabel("$x_1$", fontsize=18)
plt.ylabel("$y$", rotation=0, fontsize=18)
plt.axis([0, 2, 0, 15])

plt.show()
```

확인 및 비교

```
theta
```

```
array([[4.21076011],
       [2.74856079]])
```

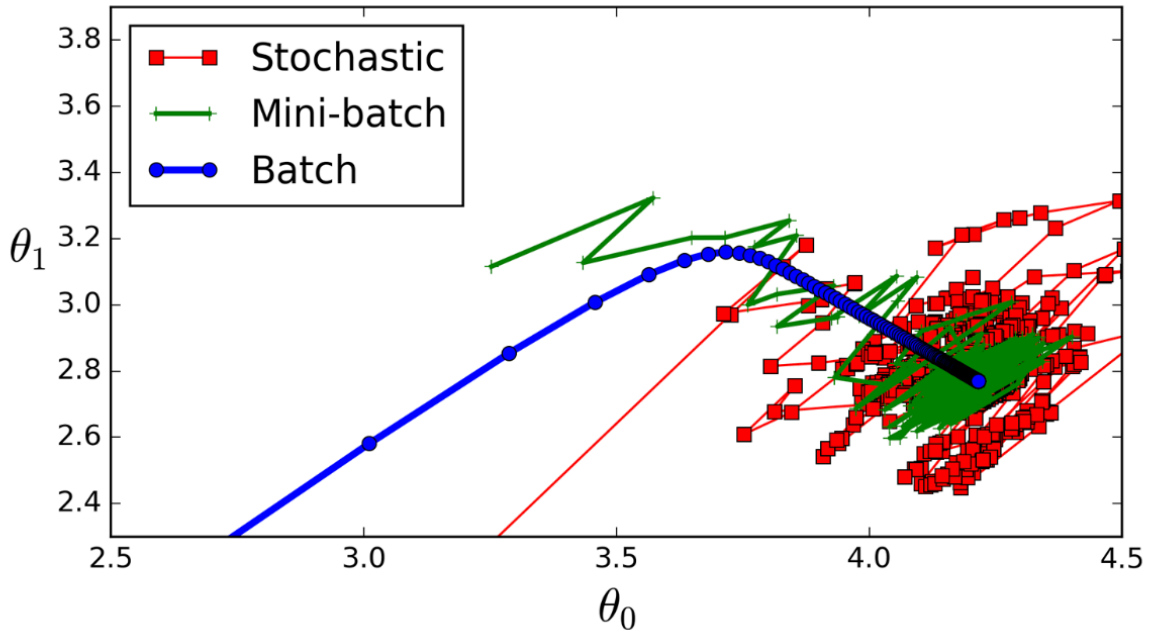
```
from sklearn.linear_model import SGDRegressor
sgd_reg = SGDRegressor(max_iter=5, penalty=None, eta0=0.1, random_state=42)
sgd_reg.fit(X, y.ravel())
```

```
sgd_reg.intercept_, sgd_reg.coef_
```

```
(array([4.10549653]), array([2.86315909]))
```

4.2.3 미니배치 경사 하강법

- 각 step에서 전체 훈련 세트(batch)나 하나의 샘플(stochastic)을 기반으로 gradient를 계산하는 것이 아니라, 임의의 작은 sample set (mini-batch)에 대해 gradient를 계산



미니배치 구현

```
theta_path_mgd = []

n_iterations = 50
minibatch_size = 20

np.random.seed(42)
theta = np.random.randn(2,1)  # 무작위 초기화

t0, t1 = 200, 1000
def learning_schedule(t):
    return t0 / (t + t1)

t = 0
for epoch in range(n_iterations):
    shuffled_indices = np.random.permutation(m)
    X_b_shuffled = X_b[shuffled_indices]
    y_shuffled = y[shuffled_indices]
    for i in range(0, m, minibatch_size):
        t += 1
        xi = X_b_shuffled[i:i+minibatch_size]
        yi = y_shuffled[i:i+minibatch_size]
        gradients = 2/minibatch_size * xi.T.dot(xi.dot(theta) - yi)
        eta = learning_schedule(t)
        theta = theta - eta * gradients
        theta_path_mgd.append(theta)
```

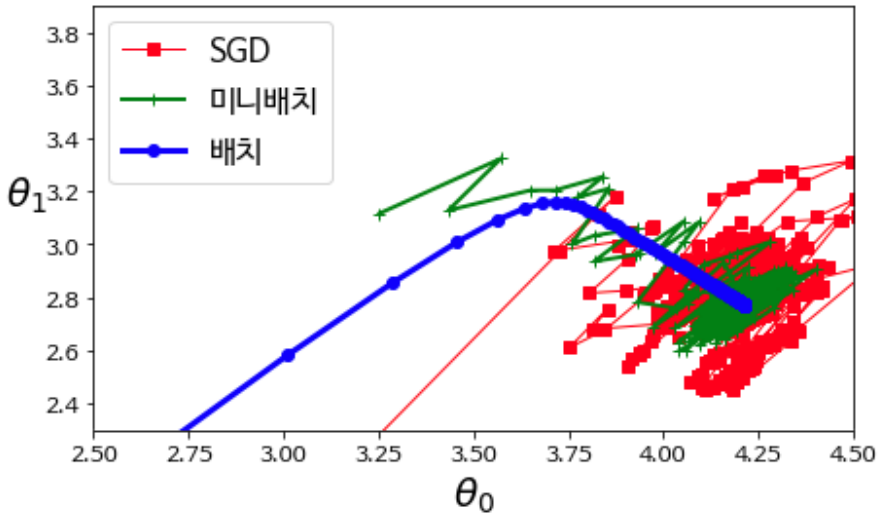
```
theta
```

```
array([[4.25214635],
       [2.7896408 ]])
```

BGD / SGD / MGD 비교

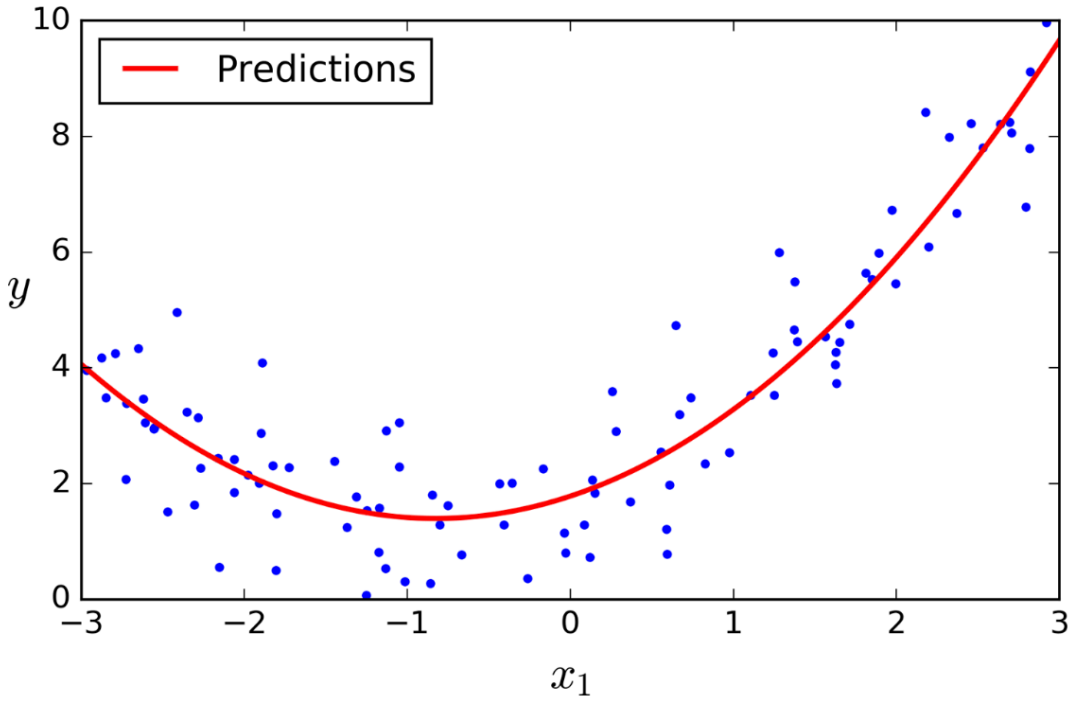
```
theta_path_bgd = np.array(theta_path_bgd)
theta_path_sgd = np.array(theta_path_sgd)
theta_path_mgd = np.array(theta_path_mgd)
```

```
plt.figure(figsize=(7,4))
plt.plot(theta_path_sgd[:, 0], theta_path_sgd[:, 1], "r-s", linewidth=1, label="SGD")
plt.plot(theta_path_mgd[:, 0], theta_path_mgd[:, 1], "g-+", linewidth=2, label="미니배치")
plt.plot(theta_path_bgd[:, 0], theta_path_bgd[:, 1], "b-o", linewidth=3, label="배치")
plt.legend(loc="upper left", fontsize=16)
plt.xlabel(r"$₩theta_0$", fontsize=20)
plt.ylabel(r"$₩theta_1$   ", fontsize=20, rotation=0)
plt.axis([2.5, 4.5, 2.3, 3.9])

plt.show()
```



4.3 다항 회귀 (Polynomial Regression)

- 비선형 데이터를 학습하는데도 선형 모델 사용 가능
- 다항 회귀(Polynomial Regression)
  - 각 특성의 거듭제곱을 새로운 특성으로 추가
  - 확장된 특성을 포함한 data set에 선형 모델을 훈련

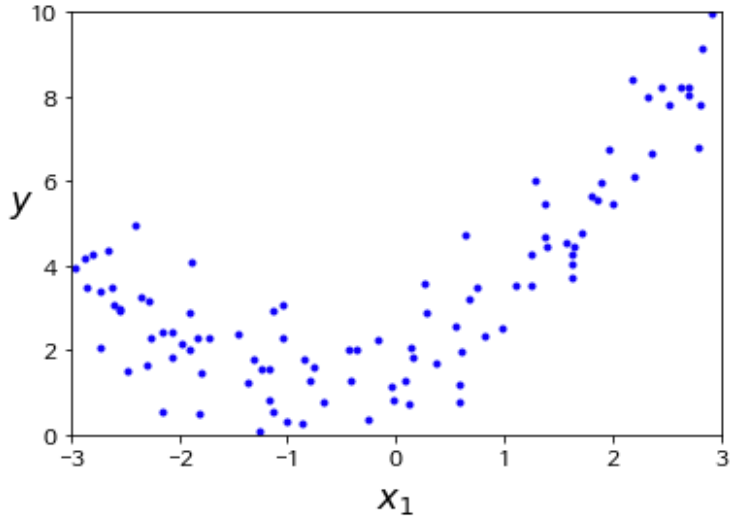


간단한 비선형 데이터 생성

```
import numpy as np
import numpy.random as rnd

np.random.seed(42)
```

```
# 간단한 2차 방정식으로 비선형 데이터 생성 (약간의 노이즈 포함)
m = 100
X = 6 * np.random.rand(m, 1) - 3
y = 0.5 * X**2 + X + 2 + np.random.randn(m, 1)
```

```
plt.plot(X, y, "b.")
plt.xlabel("$x_1$", fontsize=18)
plt.ylabel("$y$", rotation=0, fontsize=18)
plt.axis([-3, 3, 0, 10])

plt.show()
```



다항회귀 구현

◆ 훈련 데이터 변환 (새로운 특성 추가)

```
# 사이킷런의 PolynomialFeatures 사용하여 변환
from sklearn.preprocessing import PolynomialFeatures
poly_features = PolynomialFeatures(degree=2, include_bias=False)
X_poly = poly_features.fit_transform(X)
X[0]
```

```
array([-0.75275929])
```

```
X_poly[0]  # 원래 특성 X 와 특성의 제곱 (추가된 특성)
```

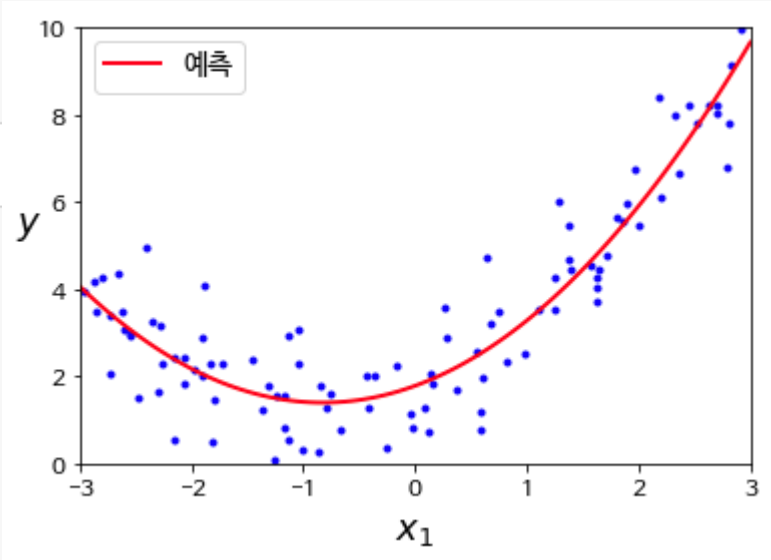
```
array([-0.75275929,  0.56664654])
```

◆ 확장된 훈련 데이터에 선형회귀 적용

```
# 확장된 훈련 데이터에 LinearRegression 적용
lin_reg = LinearRegression()
lin_reg.fit(X_poly, y)
lin_reg.intercept_, lin_reg.coef_
```

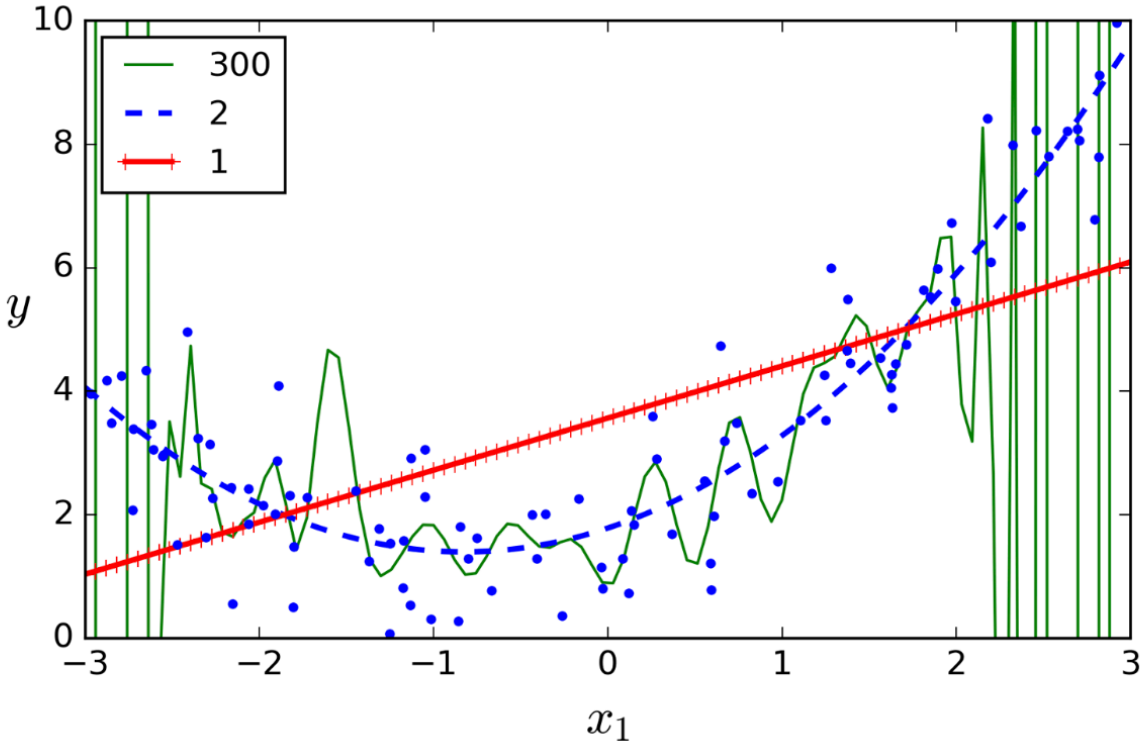
```
(array([1.78134581]), array([[0.93366893, 0.56456263]]))
```

```
X_new=np.linspace(-3, 3, 100).reshape(100, 1)
X_new_poly = poly_features.transform(X_new)
y_new = lin_reg.predict(X_new_poly)
plt.plot(X, y, "b.")
plt.plot(X_new, y_new, "r-", linewidth=2, label="예측")
plt.xlabel("$x_1$", fontsize=18)
plt.ylabel("$y$", rotation=0, fontsize=18)
plt.legend(loc="upper left", fontsize=14)
plt.axis([-3, 3, 0, 10])

plt.show()
```



다항 회귀 모델의 과대적합

- 고차 다항 회귀를 적용하면 보통 선형회귀보다 훨씬 더 training data에 잘 맞게 model을 구성하려 할 것임
- 1차, 2차, 300차 다항 회귀 모델을 이전(2차) training data에 적용시킨 결과



- 선형 모델은 underfitting, 300차 회귀 모델은 overfitting이 나타남

구현

```
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline

for style, width, degree in (("g-", 1, 300), ("b--", 2, 2), ("r-+", 2, 1)):
    polybig_features = PolynomialFeatures(degree=degree, include_bias=False)
    std_scaler = StandardScaler()
    lin_reg = LinearRegression()
    polynomial_regression = Pipeline([
            ("poly_features", polybig_features),
            ("std_scaler", std_scaler),
            ("lin_reg", lin_reg),
        ])
    polynomial_regression.fit(X, y)
    y_newbig = polynomial_regression.predict(X_new)
    plt.plot(X_new, y_newbig, style, label=str(degree), linewidth=width)

plt.plot(X, y, "b.", linewidth=3)
plt.legend(loc="upper left")
plt.xlabel("$x_1$", fontsize=18)
plt.ylabel("$y$", rotation=0, fontsize=18)
plt.axis([-3, 3, 0, 10])

plt.show()
```

4.4 학습 곡선

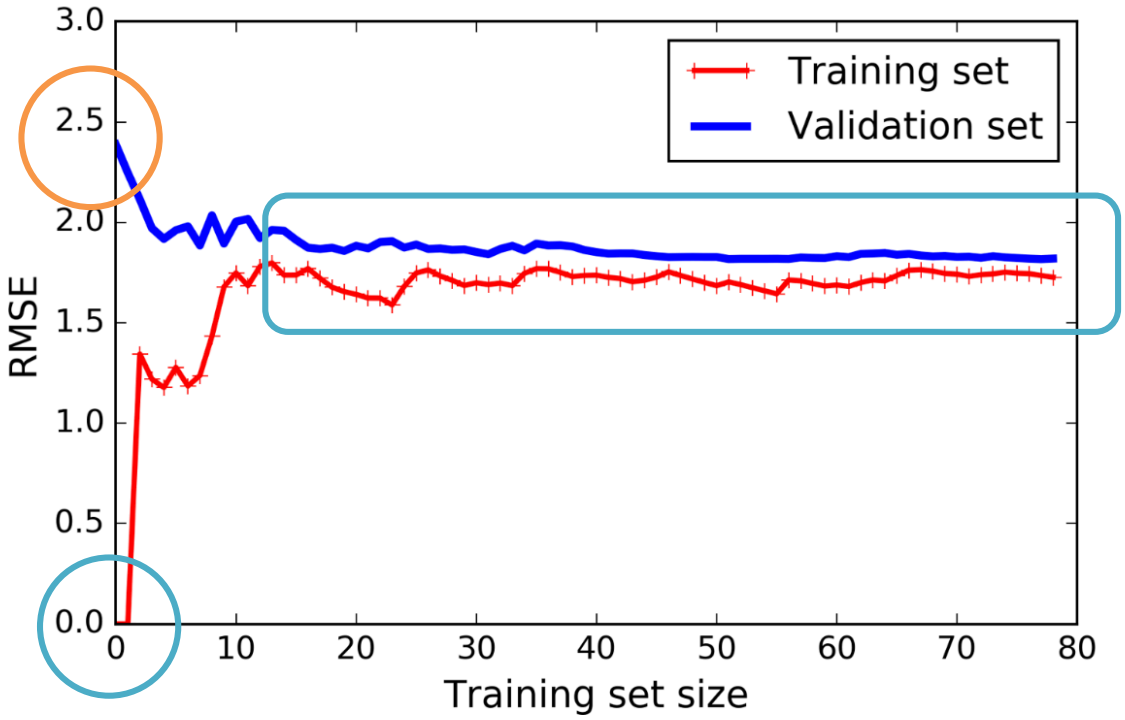
- 모델의 과대적합 / 과소적합 판단 : 교차검증 이용
- 훈련데이터에서 성능 좋음 / 교차검증 점수 나쁨 → 과대적합
- 둘 다 좋지 않음 → 과소적합

- 또 다른 방법 : 학습 곡선(그래프)
- 훈련 세트/검증 세트 모델 성능 → 훈련 세트 크기의 함수로 표현
- 훈련 세트에서 크기가 다른 서브 세트를 만들어 모델을 여러번 훈련

단순 선형 회귀 모델의 학습 곡선

- **underfitting model의 전형적인 모습**(simple linear regression)

- 두 곡선이 높은 오차에서 가까이 근접해 수평한 구간을 만든다.
- 훈련 샘플을 더 추가해도 효과 없음 → 더 복잡한 모델을 사용하거나 더 나은 특성 선택 필요



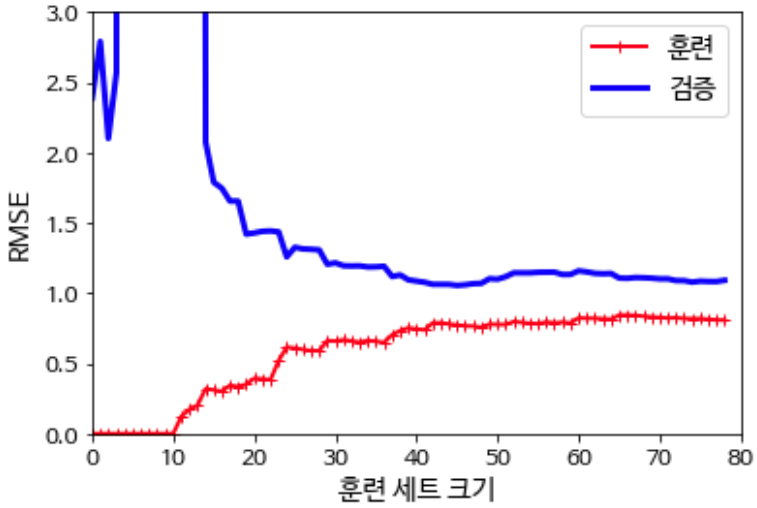
선형회귀 학습곡선 구현

```
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split

def plot_learning_curves(model, X, y):
    X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2, random_state=10)
    train_errors, val_errors = [], []
    for m in range(1, len(X_train)):
        model.fit(X_train[:m], y_train[:m])
        y_train_predict = model.predict(X_train[:m])
        y_val_predict = model.predict(X_val)
        train_errors.append(mean_squared_error(y_train[:m], y_train_predict))
        val_errors.append(mean_squared_error(y_val, y_val_predict))

    plt.plot(np.sqrt(train_errors), "r-+", linewidth=2, label="훈련")
    plt.plot(np.sqrt(val_errors), "b-", linewidth=3, label="검증")
    plt.legend(loc="upper right", fontsize=14)
    plt.xlabel("훈련 세트 크기", fontsize=14)
    plt.ylabel("RMSE", fontsize=14)
```

```
lin_reg = LinearRegression()
plot_learning_curves(lin_reg, X, y)
plt.axis([0, 80, 0, 3])

plt.show()
```

다항(10차)회귀 학습곡선 구현

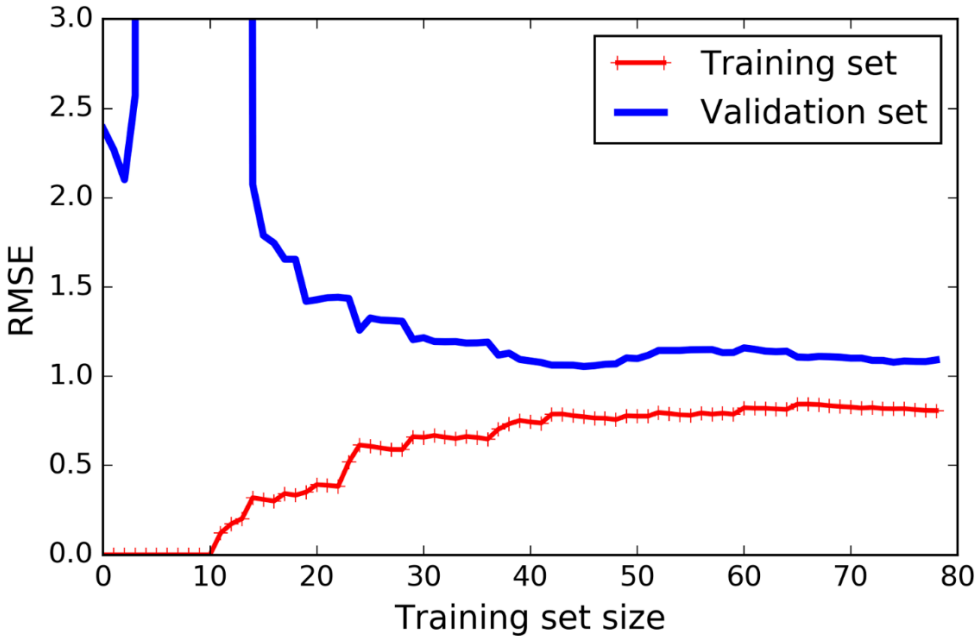
```
from sklearn.pipeline import Pipeline

polynomial_regression = Pipeline([
        ("poly_features", PolynomialFeatures(degree=10, include_bias=False)),
        ("lin_reg", LinearRegression()),
    ])

plot_learning_curves(polynomial_regression, X, y)
plt.axis([0, 80, 0, 3])

plt.show()
```



다항(10차) 회귀 모델의 학습 곡선

- training data의 오차가 앞선 일반 선형 회귀 모델에 비해 훨씬 낮음
- Training set size가 커져도 두 곡선 사이에 공간이 존재 (overfitting model의 특징)
- 더 큰 training set을 사용하면 두 곡선이 점점 가까워 짐

- **overfitting model을 개선하는 방법**
  - 검증 오차가 훈련 오차에 근접할 때 까지 더 많은 training data를 추가



편향/분산 트레이드오프

- **Bias (편향)**
  - 일반화 오차 중에서 편향은 잘못된 가정에 의해 발생
  - 예) 2차원 데이터를 선형으로 가정
  - bias가 큰 모델은 training data에 underfitting되기 쉬움

- **Variance (분산)**
  - training data에 있는 작은 변동에 모델이 과도하게 민감하기 때문에 나타남
  - 자유도가 높은 회귀 모델(ex. 고차 다항 회귀 모델)이 높은 분산을 가지기 쉬워 training data에 overfitting되는 경향

- **모델의 복잡도가 커지면 통상적으로 분산이 늘어나고 편향은 줄어든다, 반대로 모델의 복잡도가 줄어들면 편향이 커지고 분산이 작아진다.**

4.5 규제가 있는 선형 모델

- **Regularized Linear Models**
  - 과대 적합을 감소시키는 방법 : 모델 규제
  - 다항회귀 모델 규제 : 다항식의 차수를 감소 시키는 방법 사용
  - 선형회귀 모델 규제 : 모델의 가중치를 제한하는 방법 사용

  - 가중치 제한 방법 3가지 :
  - Ridge Regression, Lasso Regression, Elastic Net

4.5.1 릿지 회귀 (Ridge Regression)

- 티호노프 (러시아 수학자) 규제 (Tikhonov Regularization)

- 규제항 $\alpha \sum_{i=1}^{n} \theta_1^2$ 비용함수에 추가
- 모델의 가중치가 가능한 작게 유지되도록 함

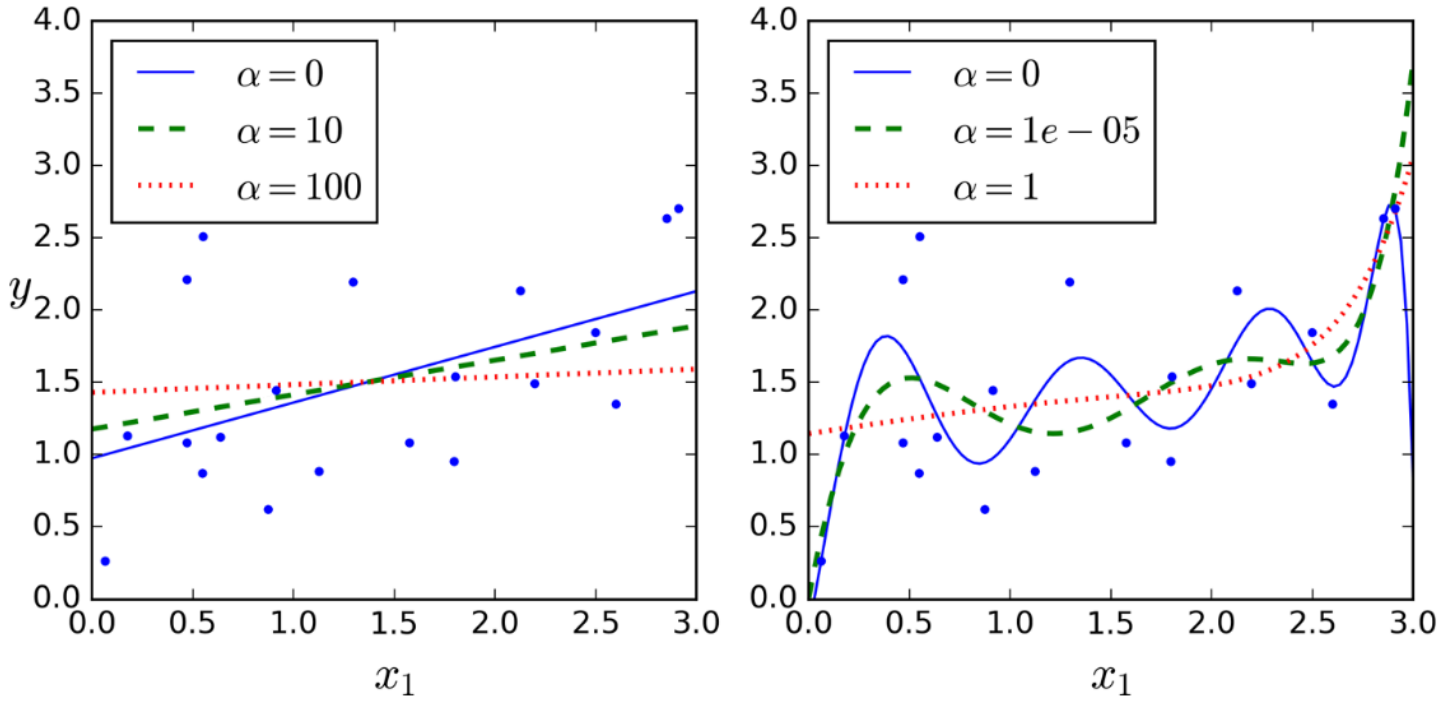
- 규제항은 훈련 기간에만 비용함수에 추가됨. (훈련이 끝나면 규제가 없는 성능 지표로 평가)
- 하이퍼파라미터인 α는 어느 정도로 모델을 규제할 지 결정
  - α=0 : 릿지회귀 = 선형회귀
  - α=매우큰값 : 모든 가중치가 거의 0에 가까워짐. 수평선

- Ridge Regression의 비용 함수

$$J(\theta) = MSE(\theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2$$

- **몇 가지 α를 사용해 릿지 모델을 훈련시킨 결과**
  - (왼쪽) 평범한 릿지 모델 (선형예측)
  - (오른쪽) 데이터 확장 → 스케일 조정 → 릿지 모델 적용 (다항회귀)

  - α값이 증가할수록 직선에 가까워짐



릿지 회귀 구현

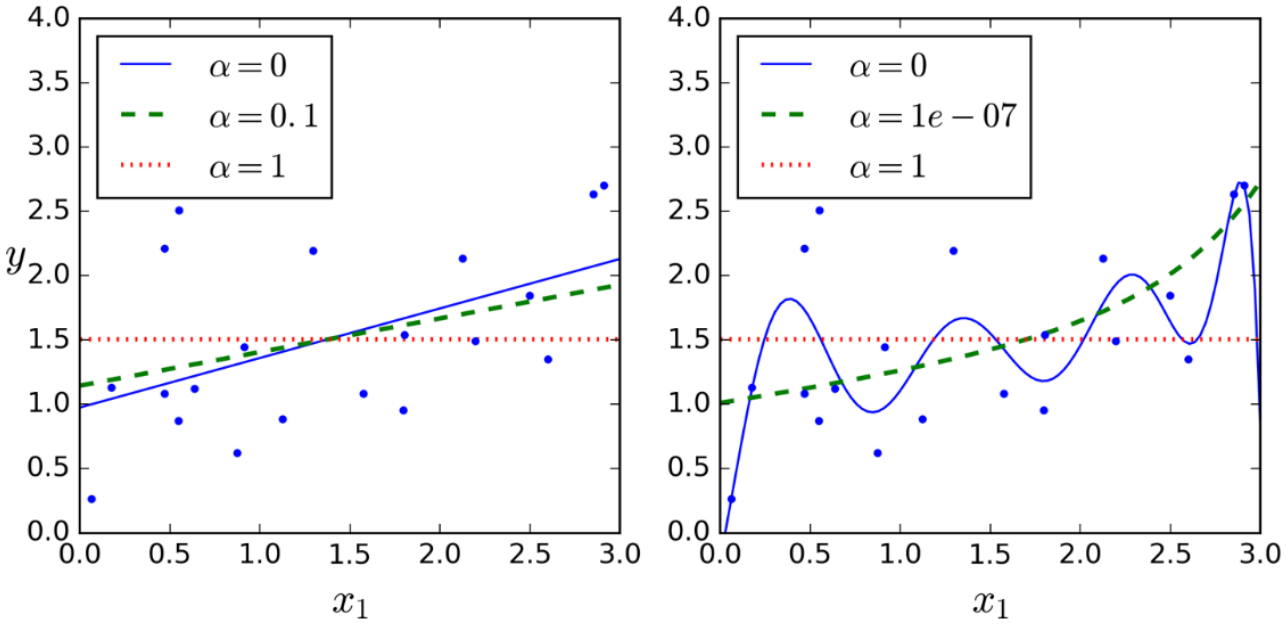
```
from sklearn.linear_model import Ridge

np.random.seed(42)
m = 20
X = 3 * np.random.rand(m, 1)
y = 1 + 0.5 * X + np.random.randn(m, 1) / 1.5
X_new = np.linspace(0, 3, 100).reshape(100, 1)

def plot_model(model_class, polynomial, alphas, **model_kargs):
    for alpha, style in zip(alphas, ("b-", "g--", "r:")):
        model = model_class(alpha, **model_kargs) if alpha > 0 else LinearRegression()
        if polynomial:
            model = Pipeline([
                    ("poly_features", PolynomialFeatures(degree=10, include_bias=False)),
                    ("std_scaler", StandardScaler()),
                    ("regul_reg", model),
                ])
        model.fit(X, y)
        y_new_regul = model.predict(X_new)
        lw = 2 if alpha > 0 else 1
        plt.plot(X_new, y_new_regul, style, linewidth=lw, label=r"$\alpha = {}$".format(alpha))
    plt.plot(X, y, "b.", linewidth=3)
    plt.legend(loc="upper left", fontsize=15)
    plt.xlabel("$x_1$", fontsize=18)
    plt.axis([0, 3, 0, 4])

plt.figure(figsize=(8,4))
plt.subplot(121)
plot_model(Ridge, polynomial=False, alphas=(0, 10, 100), random_state=42)
plt.ylabel("$y$", rotation=0, fontsize=18)
plt.subplot(122)
plot_model(Ridge, polynomial=True, alphas=(0, 10**-5, 1), random_state=42)

plt.show()
```

4.5.2 라쏘 회귀

- **Lasso(Least Absolute Shrinkage and Selection Operator) Regression**
  - 릿지 회귀와 비슷하지만 조금 다른 비용함수 사용
  - Lasso Regression 비용 함수

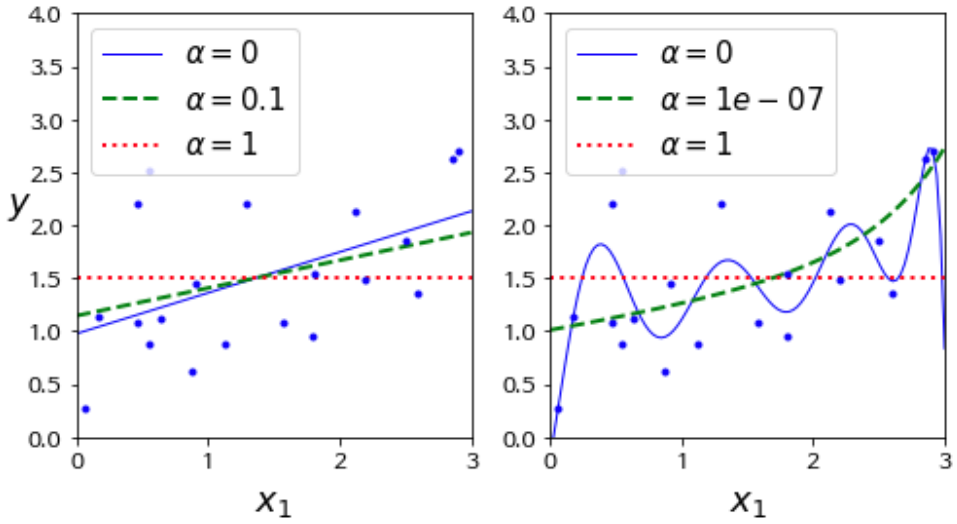
$$J(\theta) = MSE(\theta) + \alpha \sum_{i=1}^{n} |\theta_i|$$

  - 덜 중요한 가중치를 완전히 제거하려고 함



랏쏘 회귀 구현

```
from sklearn.linear_model import Lasso

plt.figure(figsize=(8,4))
plt.subplot(121)
plot_model(Lasso, polynomial=False, alphas=(0, 0.1, 1), random_state=42)
plt.ylabel("$y$", rotation=0, fontsize=18)
plt.subplot(122)
plot_model(Lasso, polynomial=True, alphas=(0, 10**-7, 1), tol=1, random_state=42)

plt.show()
```



Ridge vs. Lasso

- **예) 10,000 개의 변수를 가진 큰 data set이 존재**
- **그리고 이 변수들 중에는 서로 상관된 변수들이 존재**

  - 1) Ridge regression을 사용하면 모든 변수를 가지고 오면서 계수 값을 줄일 것이다. 하지만 문제는 1만개의 변수를 그대로 유지하므로 여전히 model이 복잡한 상태이다. 이는 모델 성능 저하에 영향을 미칠 수 있다.

  - 2) Lasso regression을 적용하면, 서로 correlate된 변수들 중에서 Lasso는 단 한개의 변수만 채택하고 다른 변수들의 계수를 0으로 바꿈. 이는 정보가 손실됨에 따라 정확성이 떨어지는 결과를 가져올 수 있다.

4.5.3 엘라스틱넷

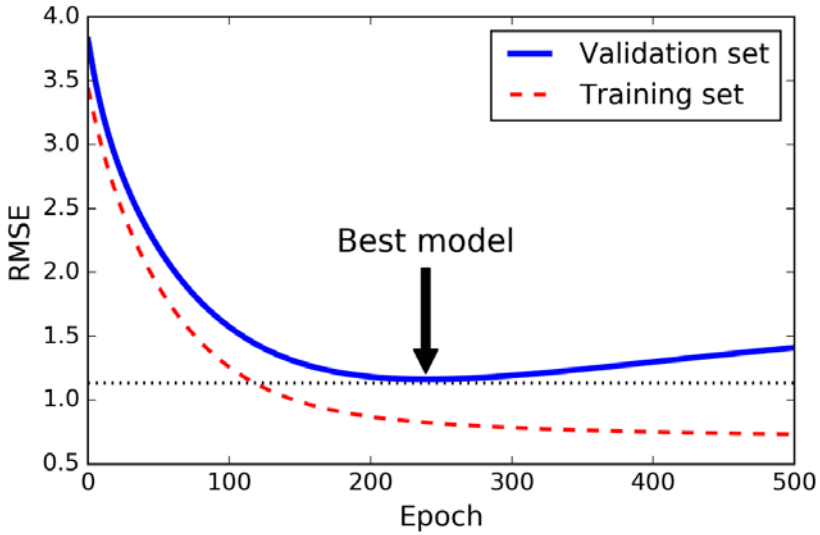
- **Elastic Net**
  - Ridge와 Lasso model을 절충한 모델
  - 규제항은 Ridge와 Lasso의 규제항을 더해서 사용
  - 혼합 비율(r)을 사용해 조절
  - r=0이면 Ridge, r=1이면 Lasso regression과 같아진다.

$$J(\theta) = \text{MSE}(\theta) + r\alpha \sum_{i=1}^{n} |\theta_i| + \frac{1-r}{2}\alpha \sum_{i=1}^{n} \theta_i^2$$

  - 대부분의 경우 **약간의 규제가 있는 model을 사용하는 것이 좋으므로** Ridge model 사용을 기본으로 하고, 실제로 쓰이는 특성이 몇 개 뿐이라고 의심되면 Lasso나 Elastic Net을 사용하는 것이 좋다.

4.5.4 조기 종료

- **Early Stopping**
  - 반복적인 학습 알고리즘을 규제하는 다른 방법
  - 검증 에러가 최솟값에 도달할 때 훈련을 중지시킴
  - 경사하강법으로 훈련시킨 고차원 다항 회귀 모델



  - Epoch 약 220정도에서 error가 가장 적게 나타나지만,
  - epoch가 증가하며 다시 error가 증가하는 overfitting 현상
  - Early stopping을 적용하면 epoch 약 220일 때, 훈련이 종료되고 최적의 파라미터를 반환

4.6 로지스틱 회귀

- **Logistic Regression**
  - sample이 특정 class에 속할 확률을 추정하는데 널리 사용
  - 추정 확률이 50%가 넘으면 모델은 sample이 해당 class에 속한다고 예측

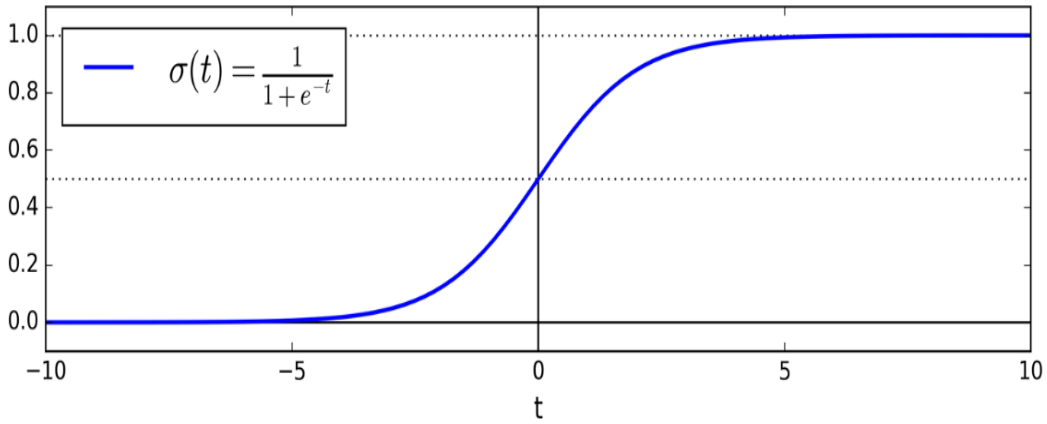
- **4.6.1 확률 추정**
  - (선형 회귀 모델처럼) 입력 특성의 가중치 합을 계산하고 bias를 더함
  - 대신 선형 회귀처럼 결과를 바로 출력하지 않고 결과값의 logistic을 출력
  - logistic regression model의 확률 추정 벡터 표현식

$$\hat{p} = h_\theta(\boldsymbol{x}) = \sigma(\theta^T \cdot \boldsymbol{x})$$

  - σ : logistic/logic이라 부르며 0과 1사이 값을 출력하는 sigmoid function

Logistic function

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$



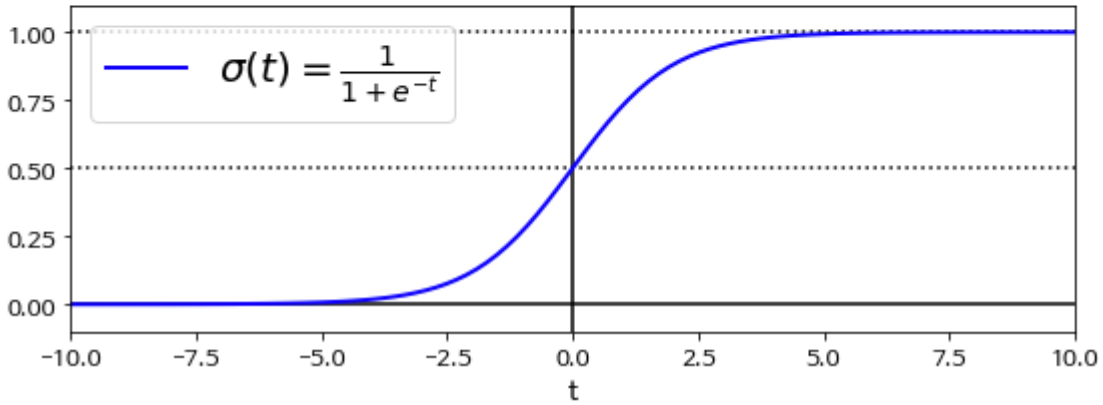
- sample x가 양성 클래스에 속할 확률 $\hat{p} = h_\theta(\boldsymbol{x})$ 를 추정하면 이에 대한 예측 $\hat{y}$ 를 쉽게 구할 수 있다.

$$\hat{y} = \begin{cases} 0 & \hat{p} < 0.5\text{일 경우} \\ 1 & \hat{p} \geq 0.5\text{일 경우} \end{cases}$$

  - t<0이면 $\sigma(t)$<0.5이고 t≥ 0이면 $\sigma(t)$ ≥0.5 이므로 logistic regression model은 $\theta^T \cdot \boldsymbol{x}$가 양수일 때 1(positive), 음수일 때 0(negative)라고 예측

로지스틱 회귀 구현

```
t = np.linspace(-10, 10, 100)
sig = 1 / (1 + np.exp(-t))
plt.figure(figsize=(9, 3))
plt.plot([-10, 10], [0, 0], "k-")
plt.plot([-10, 10], [0.5, 0.5], "k:")
plt.plot([-10, 10], [1, 1], "k:")
plt.plot([0, 0], [-1.1, 1.1], "k-")
plt.plot(t, sig, "b-", linewidth=2, label=r"$\sigma(t) = \frac{1}{1 + e^{-t}}$")
plt.xlabel("t")
plt.legend(loc="upper left", fontsize=20)
plt.axis([-10, 10, -0.1, 1.1])

plt.show()
```



4.6.2 훈련과 비용 함수

- **Logistic model의 훈련 목적**
  - 양성 샘플(y=1)에 대해서는 높은 확률을 추정,
  - 음성 샘플(y=0)에 대해서는 낮은 확률을 추정하는
  - 파라미터 벡터 θ를 찾는 것

  - 하나의 샘플에 대한 cost function

$$c(\theta) = \begin{cases} -\log(\hat{p}) & y = 1\text{일 때} \\ -\log(1-\hat{p}) & y = 0\text{일 때} \end{cases}$$

    - log function에 의해 양성 샘플을 0에 가까운 값으로 추정하면 cost가 매우 커지고, 음성 샘플을 1에 가까운 값으로 추정해도 cost가 매우 커짐

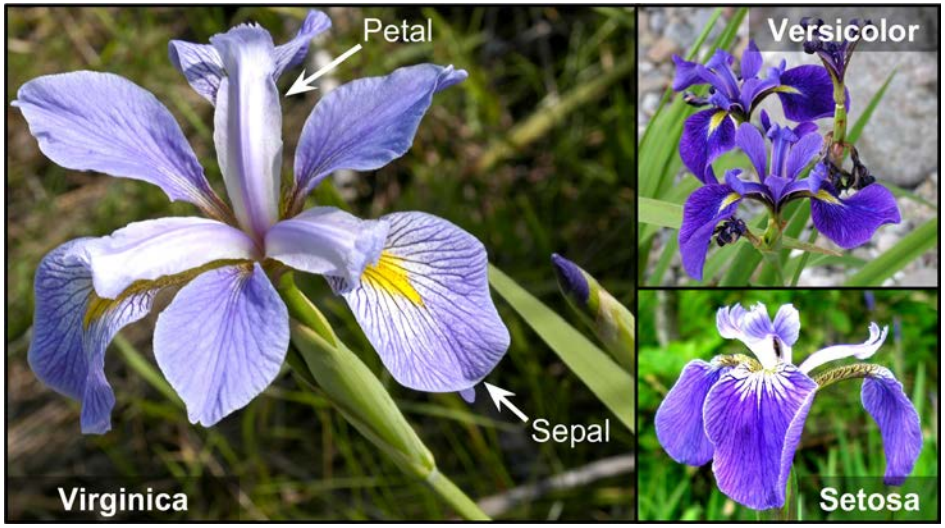
- **전체 training set에 대한 cost function**
  - 모든 훈련 샘플의 비용의 평균 : log loss

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(\hat{p}^{(i)}) + (1-y^{(i)})\log(1-\hat{p}^{(i)})]$$

  - 위 cost function의 최솟값을 계산하는 해는 없다. 다만 convex function이므로 gradient descent를 이용하면 전역 최솟값을 찾을 수 있다.

4.6.3 결정 경계

- 붓꽃 data set 분류 예
- 3개의 품종(Iris-Setosa, Iris-Versicolor, Iris-Virginica)에 속하는 붓꽃 150개의 꽃잎과 꽃받침의 너비와 길이를 포함




```
from sklearn import datasets
iris = datasets.load_iris()
list(iris.keys())
```

```
['data', 'target', 'target_names', 'DESCR', 'feature_names', 'filename']
```

```
print(iris.DESCR)
```

```
.. _iris_dataset:

Iris plants dataset
--------------------

**Data Set Characteristics:**
```

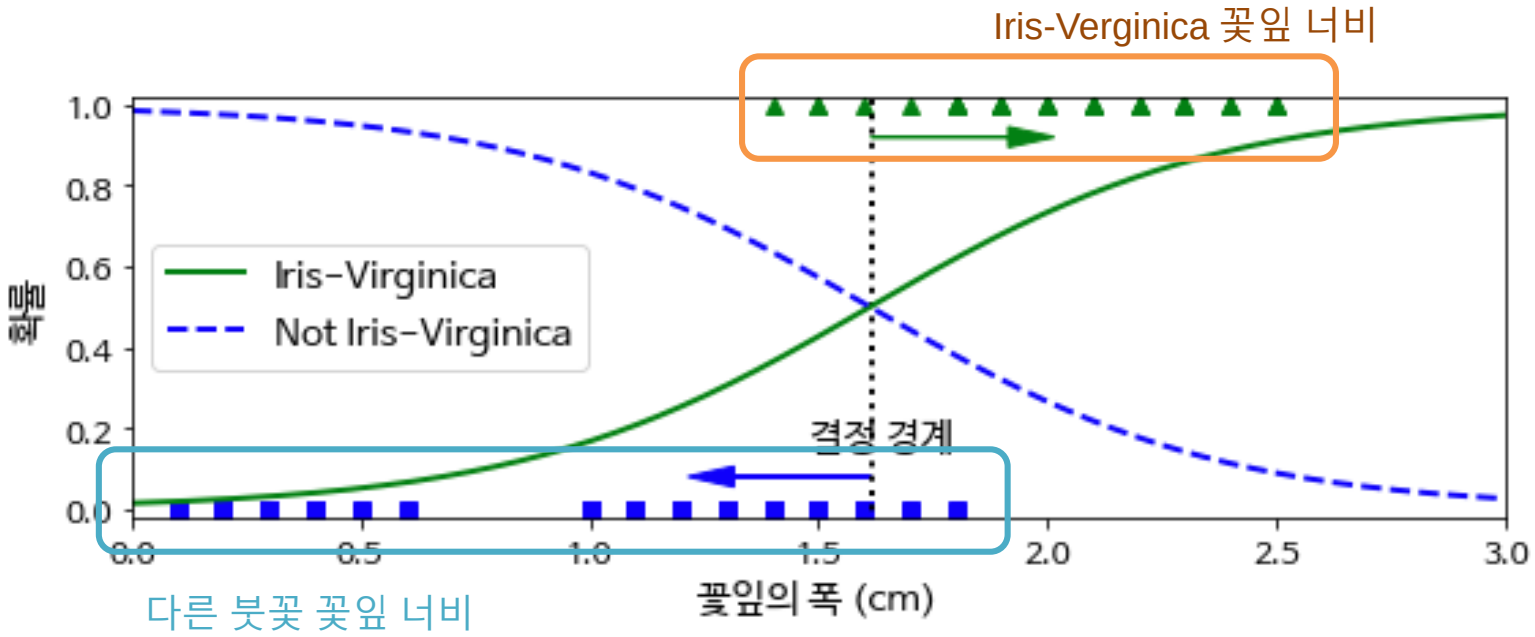
- logistic regression model을 이용해서 꽃잎의 너비가 0~3cm인 꽃에 대해 추정 확률 계산

```
X = iris["data"][:, 3:]  # 꽃잎 넓이
y = (iris["target"] == 2).astype(np.int)  # Iris-Virginica이면 1 아니면 0
```

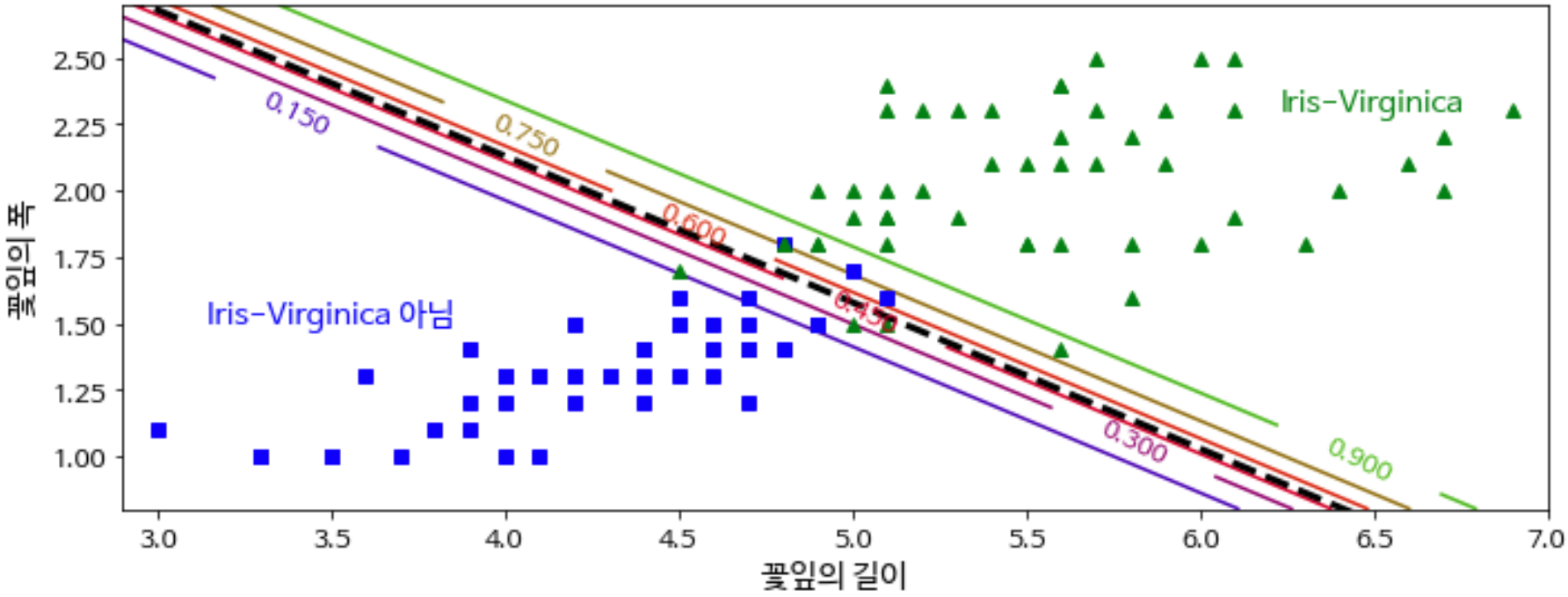
```
from sklearn.linear_model import LogisticRegression
log_reg = LogisticRegression(solver='liblinear', random_state=42)
log_reg.fit(X, y)
```

```
X_new = np.linspace(0, 3, 1000).reshape(-1, 1)
y_proba = log_reg.predict_proba(X_new)
decision_boundary = X_new[y_proba[:, 1] >= 0.5][0]

plt.figure(figsize=(8, 3))
plt.plot(X[y==0], y[y==0], "bs")
plt.plot(X[y==1], y[y==1], "g^")
plt.plot([decision_boundary, decision_boundary], [-1, 2], "k:", linewidth=2)
plt.plot(X_new, y_proba[:, 1], "g-", linewidth=2, label="Iris-Virginica")
plt.plot(X_new, y_proba[:, 0], "b--", linewidth=2, label="Not Iris-Virginica")
plt.text(decision_boundary+0.02, 0.15, "결정 경계", fontsize=14, color="k", ha="center")
plt.arrow(decision_boundary, 0.08, -0.3, 0, head_width=0.05, head_length=0.1, fc='b', ec='b')
plt.arrow(decision_boundary, 0.92, 0.3, 0, head_width=0.05, head_length=0.1, fc='g', ec='g')
plt.xlabel("꽃잎의 폭 (cm)", fontsize=14)
plt.ylabel("확률", fontsize=14)
plt.legend(loc="center left", fontsize=14)
plt.axis([0, 3, -0.02, 1.02])

plt.show()
```

추정 확률과 결정 경계


Iris-Verginica 꽃잎 너비
1.0
0.8
0.6
0.4
0.2
0.0
확률
Iris-Virginica
Not Iris-Virginica
결정 경계
0.0
0.5
1.0
1.5
2.0
2.5
3.0
꽃잎의 폭 (cm)
다른 붓꽃 꽃잎 너비

- 꽃잎 너비와 길이, 두 개의 특성 그래프



- 점선은 모델이 50% 확률을 추정하는 지점으로 이 모델의 decision boundary

선형 결정 경계 구현

```
from sklearn.linear_model import LogisticRegression

X = iris["data"][:, (2, 3)]  # petal length, petal width
y = (iris["target"] == 2).astype(np.int)

log_reg = LogisticRegression(solver='liblinear', C=10**10, random_state=42)
log_reg.fit(X, y)

x0, x1 = np.meshgrid(
        np.linspace(2.9, 7, 500).reshape(-1, 1),
        np.linspace(0.8, 2.7, 200).reshape(-1, 1),
    )
X_new = np.c_[x0.ravel(), x1.ravel()]

y_proba = log_reg.predict_proba(X_new)

plt.figure(figsize=(10, 4))
plt.plot(X[y==0, 0], X[y==0, 1], "bs")
plt.plot(X[y==1, 0], X[y==1, 1], "g^")

zz = y_proba[:, 1].reshape(x0.shape)
contour = plt.contour(x0, x1, zz, cmap=plt.cm.brg)


left_right = np.array([2.9, 7])
boundary = -(log_reg.coef_[0][0] * left_right + log_reg.intercept_[0]) / log_reg.coef_[0][1]

plt.clabel(contour, inline=1, fontsize=12)
plt.plot(left_right, boundary, "k--", linewidth=3)
plt.text(3.5, 1.5, "Iris-Virginica 아님", fontsize=14, color="b", ha="center")
plt.text(6.5, 2.3, "Iris-Virginica", fontsize=14, color="g", ha="center")
plt.xlabel("꽃잎의 길이", fontsize=14)
plt.ylabel("꽃잎의 폭", fontsize=14)
plt.axis([2.9, 7, 0.8, 2.7])

plt.show()
```

4.6.4 소프트맥스 회귀

- 직접 다중 클래스 지원하도록 일반화 : 다항 로지스틱 회귀

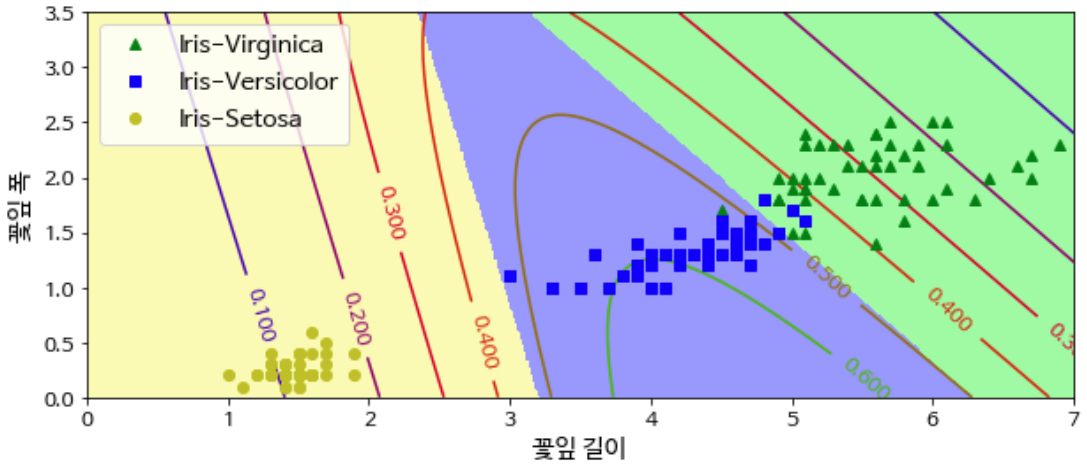
- 샘플 x를 각 클래스 k에 대한 점수 계산
- 점수에 softmax function을 적용하여 각 클래스의 확률을 추정
  - 한 번에 하나의 클래스만 예측
  - 다중 클래스, 다중 출력 안됨
    (예: 하나의 사진에서 여러 사람 얼굴 인식)

- LogisticRegression에서 ‘multi_class=multinomial’ 설정

```
X = iris["data"][:, (2, 3)]  # 꽃잎 길이, 꽃잎 넓이
y = iris["target"]

softmax_reg = LogisticRegression(multi_class="multinomial",solver="lbfgs", C=10, random_state=42)
softmax_reg.fit(X, y)
```

소프트맥스 회귀 구현

```
x0, x1 = np.meshgrid(
        np.linspace(0, 8, 500).reshape(-1, 1),
        np.linspace(0, 3.5, 200).reshape(-1, 1),
    )
X_new = np.c_[x0.ravel(), x1.ravel()]
X_new_with_bias = np.c_[np.ones([len(X_new), 1]), X_new]

logits = X_new_with_bias.dot(Theta)
Y_proba = softmax(logits)
y_predict = np.argmax(Y_proba, axis=1)

zz1 = Y_proba[:, 1].reshape(x0.shape)
zz = y_predict.reshape(x0.shape)
```




```
plt.figure(figsize=(10, 4))
plt.plot(X[y==2, 0], X[y==2, 1], "g^", label="Iris-Virginica")
plt.plot(X[y==1, 0], X[y==1, 1], "bs", label="Iris-Versicolor")
plt.plot(X[y==0, 0], X[y==0, 1], "yo", label="Iris-Setosa")

from matplotlib.colors import ListedColormap
custom_cmap = ListedColormap(['#fafab0','#9898ff','#a0faa0'])

plt.contourf(x0, x1, zz, cmap=custom_cmap)
contour = plt.contour(x0, x1, zz1, cmap=plt.cm.brg)
plt.clabel(contour, inline=1, fontsize=12)
plt.xlabel("꽃잎 길이", fontsize=14)
plt.ylabel("꽃잎 폭", fontsize=14)
plt.legend(loc="upper left", fontsize=14)
plt.axis([0, 7, 0, 3.5])
plt.show()
```

Any Questions...
Just Ask!