Introduction to Machine Learning

2016.08.

오혜연

alice.oh@kaist.edu

Slides Acknowledgements

- Based on Andrew Ng’s course
- Translated / added by 가재령

머신 러닝이란?

- **명시적으로 프로그래밍을 하지 않고도 컴퓨터가 학습할 수 있는 능력을 갖게 하는 것**

- 사용 예
  - 데이터마이닝
    - 클릭 기록, 의료 기록, 유전자 분석 등
  - 수작업으로 프로그래밍할 수 없는 것들
    - 자율 운행 헬리콥터, 얼굴 인식, 스팸 필터 등
  - 개개인의 유저에게 최적화된 추천 알고리즘
    - 아마존(상품 추천), 넷플릭스(영화 추천)

지도 학습

지도 학습 - 회귀 문제 1

- 주택 가격 데이터를 가지고 있다고 하자. 이 데이터를 바탕으로 하여 75 제곱미터 크기의 주택을 팔려고 한다면, 얼마의 가격을 받아야 할까?

지도 학습 - 회귀 문제 2

- 데이터를 지나는 직선을 긋고 그 직선을 참고하면, 1.6억 정도의 가격이 적절할 것이라는 것을 알 수 있음
- 하지만 머신 러닝은 단순히 직선을 그리는 것 이상의 것

지도 학습 - 회귀 문제 3

- 직선 대신 2차 곡선을 그리면, 집의 적정 가격은 2.2억
- 직선과 2차 곡선 중 어떤 것을 선택해야 할 지에 대한 정답은 없음
- 두 가지 방법 모두 지도 학습의 좋은 예시임

지도 학습 - 회귀 문제 4

- '지도 학습' (Supervised Learning)이란 우리가 알고리즘에게 정확한 답을 알고 있는 데이터 세트를 준다는 데에서 유래
- 이 예시에서는 실제로 집이 팔린 가격을 담은 데이터 세트를 제시하면, 알고리즘은 이를 분석하여 집을 팔아야 하는 가격을 제시함
- 좀 더 정확한 용어: **연속적인 값을 예측하는 문제인 '회귀 문제'(regression problem)**

지도 학습 - 분류 문제 1

- 분류 문제는 입력 데이터가 어떠한 (이산적인) 분류에 해당할 지에 대한 확률을 예측하는 것
- 분류는 두 개, 혹은 여러 개의 클라스

지도 학습 - 분류 문제 2

- 1개 이상의 특징/속성이 주어진 문제
  - 환자의 나이, 종양 크기, 두꺼움의 정도, 세포 크기의 균일함 등
- 분홍색 점이 양성 종양일지?

- 분류 알고리즘의 분류선 (classification boundary) 검정색 직선
- 이 종양이 boundary의 왼쪽에 있기 때문에 무해한 종양이라고 판단

비지도 학습

비지도 학습 1

- ‘레이블이 없는’ 데이터 (오른쪽 그림)

비지도 학습 2

- **군집화 알고리즘(Clustering algorithm)**
- 예: 뉴스 사이트에서의 기사 그룹을 분류

비지도 학습 4

- 비지도 학습은 이외에도 다양한 분야에 활용된다.

기계학습 알고리즘의 정확도 측정

- Supervised classification: 보지 않은 (unseen) 데이타에 대한 정확도 (accuracy)
- Unsupervised classification: 군집화 된 데이타에 대한 검증
  - Internal evaluation: 같은 클러스터에 있는 데이타의 유사도는 높고, 다른 클러스터에 있는 데이타의 유사도는 낮은지에 대한 검증:
  - External evaluation: 골드 스탠다드(정답)이라고 부를 수 있는 외부의 라벨을 이용하여 클러스터를 검증

Training, validation, test data

- Training data: 모델 학습에 사용
- Validation data: 학습을 할때 튜닝이 필요한 parameter들을 최적화하기 위해 사용
- Test data: 학습된 모델을 검증하기 위해 사용

Bias와 Variance에서 나오는 에러

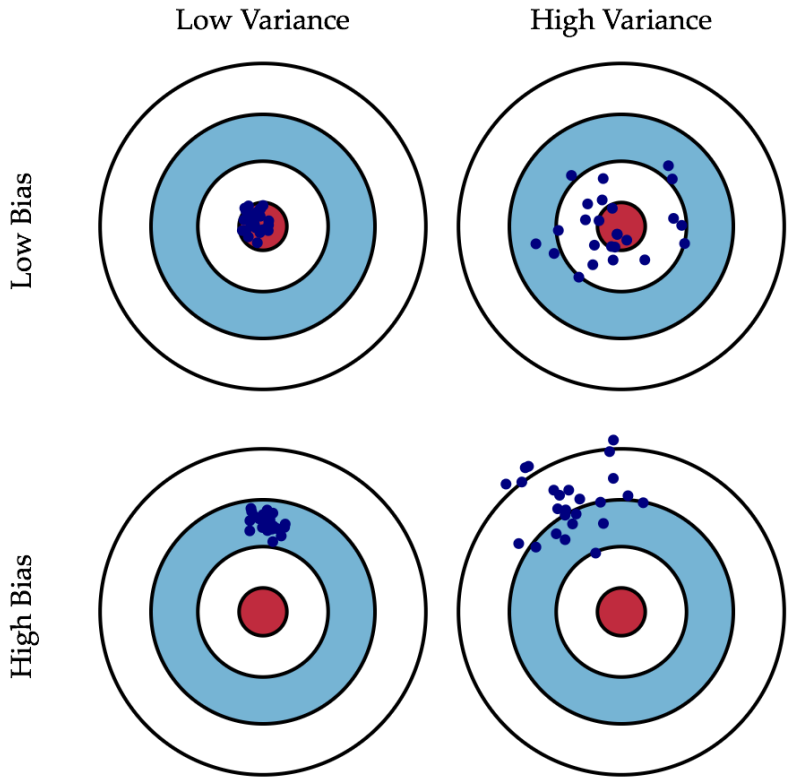


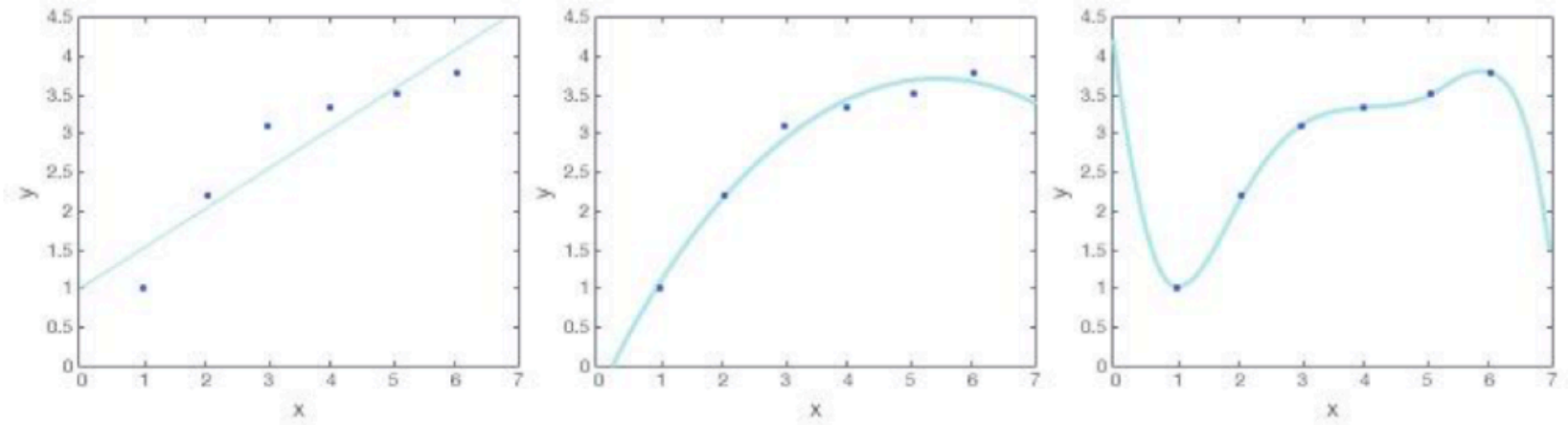
Fig. 1 Graphical illustration of bias and variance.

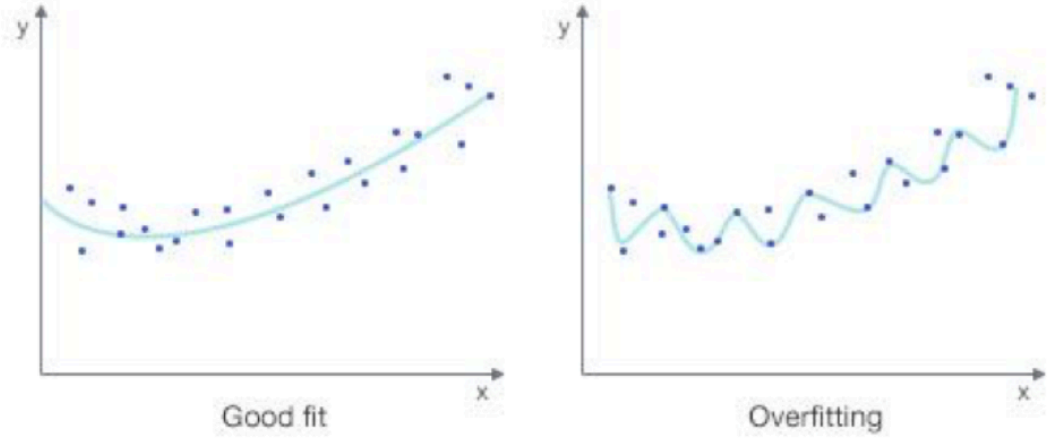
Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Bias vs Variance

- 모델이 복잡하면 bias가 낮고 variance가 높음
  - Training data에 대한 정확도 올라감
  - 대신 Test data에 대한 정확도는 내려갈 수 있음 (overfitting)
- 모델이 간단하면 bias가 높고 variance가 낮음
  - Training data에 대한 정확도 내려감
  - 대신 overfitting 문제는 덜 할 수 있음

Overfitting





Source: http://blog.fliptop.com/blog/2015/03/02/bias-variance-and-overfitting-machine-learning-overview/

모델의 복잡도와 Overfitting

Source: http://blog.fliptop.com/blog/2015/03/02/bias-variance-and-overfitting-machine-learning-overview/

주택 가격 예측 문제 - 재고

- 예전 슬라이드로 돌아가서 주택 가격 예측 문제를 재고해 보자.
- 이 문제는 지도 학습이고, 회귀 문제이다.

표기법

m = 훈련용 데이터 세트의 개수

x = 입력 변수/특징

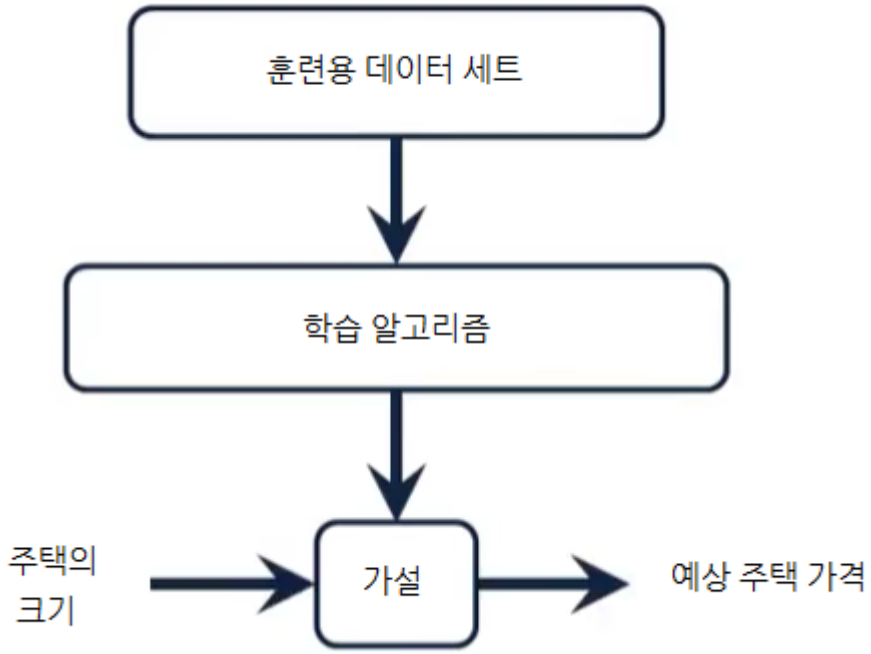
y = 출력 변수/특징

(x, y) : 한 훈련 데이터

$(x^{(i)}, y^{(i)})$: i번째 훈련 데이터

가설

- 학습 알고리즘이 훈련용 데이터 세트를 학습하면, ‘주택의 크기’ (x)를 받아 ‘예상 주택 가격’ (y)를 결과값으로 내보내는 ‘가설’을 만들게 된다.



가설을 어떻게 묘사할 것인가?
- 단일 변수 선형 회귀 1

- 가설을 묘사하는 방법에는 여러 가지가 있겠지만, 가장 간단한 1차 함수부터 시작해 보자. 아래 그림은 변수가 1개인 선형 회귀(linear regression) 그래프다.

가설을 어떻게 묘사할 것인가?
- 단일 변수 선형 회귀 2

- 주어진 데이터 세트에서 매개 변수를 어떻게 설정해야 할까?

가설을 어떻게 묘사할 것인가?
- 단일 변수 선형 회귀 3

- 매개 변수에 따라 가설의 함수의 모양이 달라질 것이다. h(x)의 값이 y의 값에 가깝도록 매개 변수를 설정하면 될 것이다.

가설을 어떻게 묘사할 것인가?
- 단일 변수 선형 회귀 4

- $$J(\theta_0,\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

- 의 식이 최소가 되게 하는 매개 변수가 이에 대한 답일 것이다.
- 이 때 J를 '비용 함수' 라고 부른다.

가설을 어떻게 묘사할 것인가?
- 단일 변수 선형 회귀 5

- 이를 그래프로 나타내면 다음과 같다.

- 여기에서 비용 함수가 최소가 되는 점을 찾기 위해서는 등고선을 그리면 좋다.

가설을 어떻게 묘사할 것인가?
- 단일 변수 선형 회귀 6

- 이 점에 해당하는 매개변수값이 왼쪽의 그래프에 그려졌다.
- 훈련 데이터에 대한 비용 h(x)의 값과 실제 데이터의 차이가 적어질수록 J값이 감소하므로 현재의 가설이 정확해질수록 등고선 안쪽으로 두 번째 그래프의 점이 점점 가까워진다.

구배법 (gradient descent) 1

- 그래서, 우리는 이 '비용 함수' 즉 J값을 줄이기 위해 머신 러닝을 적용하고자 한다.
- 어떤 매개변수 값에서 시작하여 이 매개변수 값을 조정해 나가면서, 비용 함수인 J값이 최소가 되는 점을 찾는 것이 이 알고리즘의 개괄이다.
- 이 점에서 시작하여 우리가 원하는 비용 함수가 최소가 되는 점을 찾으려고 한다.
  어떻게 해야 할까?

구배법 (gradient descent) 2

- 이 그래프를 하나의 언덕이라고 가정하자. 여기서 우리가 해야 할 일은 360도를 돌아보면서, 어떤 방향으로 한 걸음을 내딛어야 언덕을 가장 빠르게 내려갈 수 있는지에 대한 답을 찾는 것이다.
- 각각의 지점에서 '가장 빠르게 내려갈 수 있는 방향'으로 한 걸음씩 내딛다 보면, 우리는 극솟값(local minimum)에 도달할 것이다.

구배법 (gradient descent) 3

- 구배법은 흥미로운 특징을 하나 가지고 있다. 우리가 이 극솟값을 찾을 때의 시작 지점과 다른 곳에서 시작할 경우, 다른 극솟값이 나올 수도 있다는 것이다. 예를 들어, 언덕의 예전 시작 지점으로부터 세타 1의 음의 방향으로 약간만 떨어진 곳에서 시작한다면, 아래 그림처럼 전혀 다른 지점에 도착하게 된다.

구배법 4 - 알고리즘

- 다음은 구배법의 알고리즘이다.

수렴할 때 까지 다음을 반복

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad (\text{for } j = 0 \text{ and } j = 1)$$

- 이 알고리즘에서 주의해야 할 점은 각 매개변수의 값이 동시에 갱신되어야 한다는 것이다.

구배법 4 - 알고리즘

- 여기서 알파(α)는 학습률(learning rate)로, '한 발자국'의 크기가 얼마만큼일 것인지를 결정한다. 이 값이 너무 크면 원하는 값을 찾기도 전에 다른 방향으로 가게 되어 원하는 값을 찾지 못할 수 있고, 이 값이 너무 작으면 매개변수의 변화량이 작아져 원하는 값을 찾는데 오랜 시간이 소요될 것이다.

구배법 4 - 알고리즘

- 학습률이 고정된 값이라도, 극솟값에 다가가면서 구배법은 자연스럽게 더 '조금씩' 움직이게 된다. 따라서 시간이 지나더라도 학습률을 감소시킬 필요는 없게 된다.

구배법 5
알고리즘: 선형 재귀 모형에서의 사례

- 이제 구배법의 알고리즘을 선형 재귀 모형에 적용해 보자.

수렴할 때 까지 다음을 반복

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad (\text{for } j = 0 \text{ and } j = 1)$$

$$h_\theta(x) = \theta_0 + \theta_1 x$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

- 수학적 계산을 통해 이 알고리즘은 이 식으로 바뀐다.

수렴할 때까지 다음을 반복

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \cdot x^{(i)}$$

구배법 6
알고리즘: 선형 재귀 모형에서의 사례

- 선형 재귀 모형에서는 앞서 살펴본 경우와 달리 여러 개의 극솟값을 가지지 않고 단 하나의 극솟값(=최솟값)을 가진다. 따라서 어떤 점에서 시작하더라도 항상 하나의 점으로 수렴하게 된다.

주택 가격 문제 - 재고

- 우리는 지난 슬라이드에서 주택 가격을 결정하는 요인이 사이즈 하나일 때의 주택 가격을 예상하는 방법을 알아보았다. 하지만 만약, 주택 가격을 결정하는 요인이 한 개가 아니면 가격을 어떻게 예상할 수 있을까?

사이즈	방의 수	층의 수	주택의 나이	가격
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
…	…	…	…	…

주택 가격 문제 - 표기법

n = 특징의 수

$x^{(i)}$ = i번째 데이터

$x_j^{(i)}$ = i번째 데이터의 j번째 특징의 값

사이즈	방의 수	층의 수	주택의 나이	가격
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...

주택 가격 문제

- 위 데이터는 4개의 특징(사이즈, 방의 수, 층의 수, 주택의 나이)을 가지고 있으므로,

$$h_\theta(x) = \theta_0 + \theta_1 x$$

- 의 식 대신

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

- 의 식을 사용하면 될 것이다.

주택 가격 문제

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

- 이제 이 식을 벡터 형식으로 표현해 보자.

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \cdots \\ x_n \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \cdots \\ \theta_n \end{bmatrix}$$

$$h_\theta(x) = \theta^T x$$

- 이를 **다변수 선형 회귀** 라고 한다.

다변수 선형 회귀

다변수에 대한 구배법

다변수에 대한 구배법

- 일변수에 대한 구배법과 비슷하다.
- 가설 : $h_\theta(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$
- 매개 변수 : θ (n+1 차원 벡터)
- 비용 함수 : $J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$
- 알고리즘 : 반복 $\{$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

$\}$

다변수에 대한 구배법의 알고리즘

- 알고리즘을 실행하기 위해 $J(\theta)$ 를 편미분하면,

반복 $\{$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

$\}$

반복 $\{$

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

$\}$

다변수에 대한 구배법의 알고리즘

- 일변수 구배법에서도 언급했지만, θ_j 의 갱신은 동시에 이루어져야 한다.

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_1^{(i)}$$

$$\theta_2 := \theta_2 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_2^{(i)}$$

$$\cdots$$

특징 크기 조정과 평균 표준화

특징 크기 조정

- 특징 크기 조정은 각 특징들이 비슷한 규모를 가지도록 조정하는 것이다.
- 조정의 목표는 모든 특징들이 대략적으로 $-1 \leq x_i \leq 1$ 범위에 들게 하는 것이다.
- 예를 들어, 방 사이즈는 0-2000 제곱 피트, 방의 수는 1-5라고 하자. 그렇다면 방 사이즈에는 2000을 나누고, 방의 수에는 5를 나누면 될 것이다.

평균 표준화

- 평균 표준화는 특징들이 0의 평균을 가지도록 하는 것을 목표로 한다. 이 때 각각의 특성들이 $-0.5 \leq x_1 \leq 0.5, -0.5 \leq x_2 \leq 0.5$ 의 범위에 들도록 조정한다.
- 즉, $x_1 := \frac{x_1 - avg}{range(\max - \min)}$ 식을 따른다.
- 예를 들어서, 지난 예시에서 사이즈의 평균이 1000, 방의 평균이 2라면,

$$x_1 := \frac{x_1 - 1000}{2000} \qquad x_2 := \frac{x_2 - 2}{4}$$

- 가 된다.

구배법과 학습률

구배법과 학습률

- 어떻게 학습률 α 를 골라야 구배법이 올바르게 작동할까?
- 학습률이 지나치게 낮으면, 수렴이 너무 느릴 것이다. 충분히 작은 학습률에 대해서는 비용 함수가 매 반복마다 감소한다.
- 학습률이 지나치게 높으면, 비용 함수가 매 반복마다 감소하지 않거나, 아예 수렴하지 않거나, 오히려 더 느리게 수렴할 수 있을 것이다.

- 학습률α를 고르기 위해서는, 0.001부터 3배씩 숫자를 올려 가며 시도해 보는 것이 좋다. 그렇다면, 어떤 학습률에 있어서 구배법이 올바르게 작동하고 있는지를 어떻게 알 수 있을까?

구배법과 학습률

- 비용 함수가 매 반복마다 감소하고,
  그 변화 폭이 10^{-3} 보다 작으면
  비용 함수가 **수렴한다**고 판단한다.

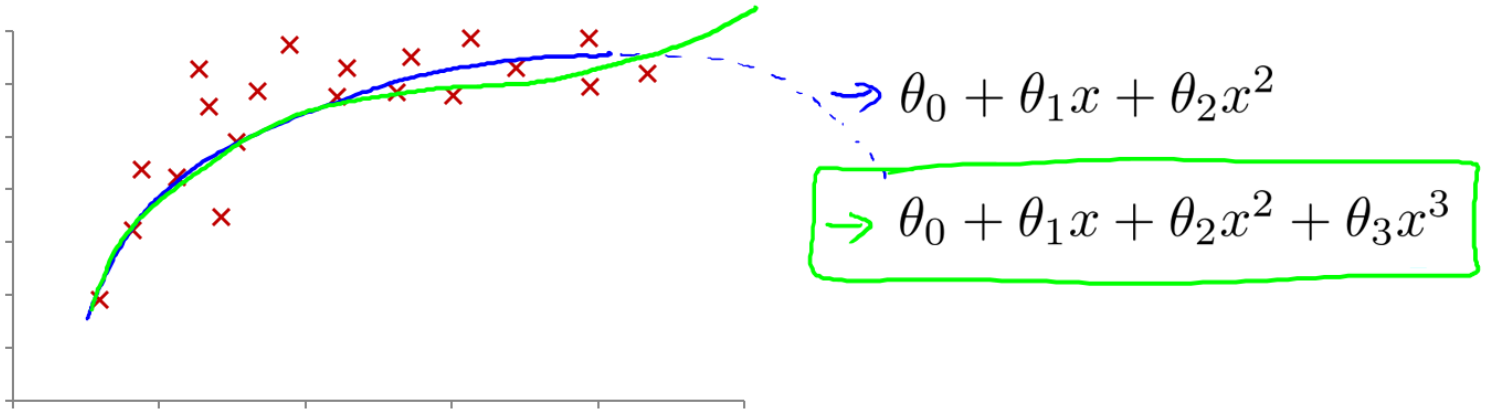
구배법과 학습률

- 비용 함수가 매 반복마다 감소하지 않으면 더 작은 학습률을 찾아야 한다.

다항 회귀

주택 가격 문제 - 다항 회귀

- 우리는 지금까지 주택 가격 문제에서 선형 회귀 사례만을 다루었다. 이 파트에서는 다항 회귀 사례를 다루고자 한다.
- Y축이 가격이고, X축이 사이즈인 다음과 같은 주택 가격 데이터가 있다.
- 선형 회귀 대신 **다항식**을 사용한 다항 회귀를 사용한다면, 파란 색 곡선과 연두색 곡선에 대하여 다음과 같은 수식을 쓸 수 있을 것이다.



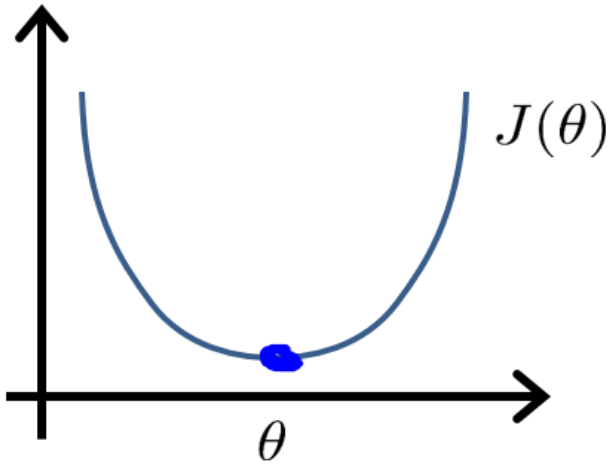
주택 가격 문제 - 다항 회귀

- 이 때 $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$
  $= \theta_0 + \theta_1 (size) + \theta_2 (size)^2 + \theta_3 (size)^3$
- 즉, $x_1 = (size)$
  $x_2 = (size)^2$
  $x_3 = (size)^3$ 이다.

- 이와 반대로, $h_\theta(x) = \theta_0 + \theta_1 (size) + \theta_2 \sqrt{(size)}$ 형태도 가능할 것이다.

표준 방정식

표준 방정식

- 구배법을 꼭 이용하지 않고서도 선형 회귀 문제를 풀 수 있다. 새로 소개할 방법인 표준 방정식을 이용한 해법은, 분석적으로(analytically) θ 에 대해 방정식을 푸는 것이다.
- $(\theta \in \mathbb{R})$ 이라고 하자. 그러면 비용 함수의 그래프는 다음과 같이 그려질 것이다.
- 

  $J(\theta) = a\theta^2 + b\theta + c$

  이 때, $\frac{d}{d\theta} J(\theta) = 0$ 이 되는 θ 가 우리가 찾고자 하는 답일 것이다.

표준 방정식 - 일반화

- 일반적인 사례에서 $\theta \in \mathbb{R}^{n+1}$ 일 때,
  모든 j에 대해 $\frac{\partial}{\partial \theta_j} J(\theta) = \cdots = 0$를 만족시키는 $\theta_0, \theta_1, \ldots, \theta_n$을 찾으면 된다.

표준 방정식

- 이를 행렬을 이용해서 풀 수 있는 방법이 있다.

x_0	사이즈 x_1	방의 수 x_2	층의 수 x_3	건물의 나이 x_4	가격 y
1	2104	5	1	45	460
1	1416	3	2	40	232
1	1534	3	2	30	315
1	852	2	1	36	178
1	3000	4	1	38	540

$$X = \begin{bmatrix} 1 & 2104 & 5 & 1 & 45 \\ 1 & 1416 & 3 & 2 & 40 \\ 1 & 1534 & 3 & 2 & 30 \\ 1 & 852 & 2 & 1 & 36 \\ 1 & 3000 & 4 & 1 & 38 \end{bmatrix} \qquad y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \\ 540 \end{bmatrix}$$

$$\Theta = (X^T X)^{-1} X^T y$$

구배법과 표준 방정식의 장단점

- 구배법은 학습률 α를 골라야 하고, 많은 반복이 필요한 단점이 있지만 특징의 수가 많을 때도 잘 동작한다.
- 표준 방정식은 학습률을 고를 필요가 없고, 반복을 할 필요가 없는 장점이 있지만 $(X^T X)^{-1}$ 을 계산하는 과정에서 O(n^3)의 복잡도가 소요되므로 특징의 수가 많을 때 상당히 느리다는 단점이 있다.

$(X^T X)$의 역행렬이 없을 때?

- (1) 일차종속(linearly dependent)인 특징을 제거한다.
  - 예 : 사이즈(제곱미터)와 사이즈(평) 중 하나를 제거

- (2) 특징의 수가 너무 많은 경우 (훈련용 데이터의 수보다 특징이 같거나 많음)
  - 몇 특징을 제거하거나, 뒤에서 배울 정규화(regularization)를 이용한다.

기계학습 강의 3:
확률론과 나이브베이즈 분류법

August 2016

Alice Oh
alice.oh@kaist.edu

이산형 랜덤 변수 (discrete random variable)

- 이산형 랜덤 변수 x 는 무제한(혹은 제한된) 값 중 하나를 갖게 된다 (예: 주사위)
- p(X = x)를 줄여서 표기하는 방법인 p(x)는 X가 어떠한 특정 값 x를 갖게 될 확률을 나타냄 (예: 주사위에 대해 p(1) = 0.166666)

이산형 랜덤 변수

- 이 확률은 무엇일까요? (A는 k개의 다른 값을 가질 수 있는 이산형 랜덤 변수임)

$$P(A = v_i \wedge A = v_j) \quad \text{when} \quad i \neq j$$

그렇다면 이것은?

$$P(A = v_i \vee A = v_2 \vee \ldots \vee A = v_k)$$

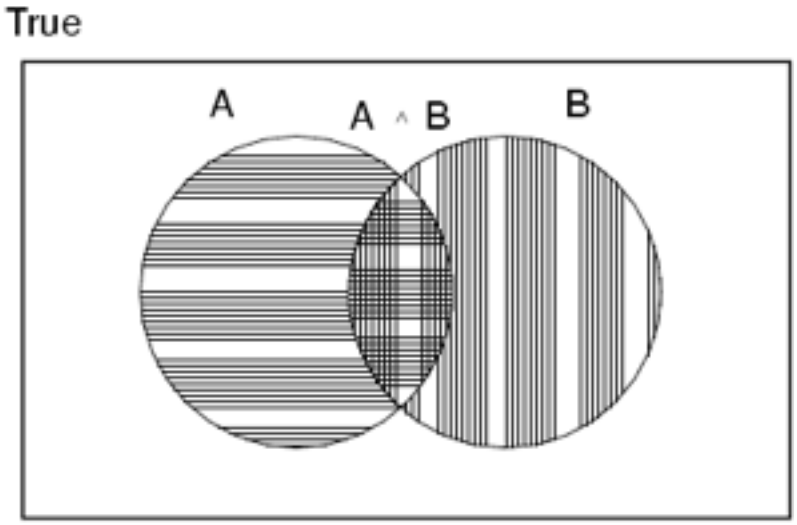
Joint 확률

- x는 43세이고, 장동건, 강동원의 팬이다.
- x가 교수일 확률은?
- x가 여성 교수일 확률은?

Joint 확률

- X는 43세이고, 장동건, 강동원의 팬이다.
- X가 교수일 확률은?
- X가 여성 교수일 확률은?

- Events A (교수), B (여성)
- P(A)? P(A^B)?
- P(A^B)는 P(A)보다 항상 크거나 같다



이산형 랜덤 변수

- 그렇다면 이것은?

$$P(B \wedge [A = v_1 \vee A = v_2 \vee ... \vee A = v_m])$$

when $m \leq k$?

이산형 랜덤 변수

$$P(B \wedge [A = v_1 \vee A = v_2 \vee ... \vee A = v_m])$$

$$= \sum_{j=1}^{m} P(B \wedge A = v_j)$$

베이즈 법칙

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

$$= \frac{P(X = x)P(Y = y|X = x)}{\sum_{x'} P(X = x')P(Y = y|X = x')}$$

베이즈 법칙

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

$$= \frac{P(Y = y)P(X = x|Y = y)}{\sum_{y'} P(Y = y')P(X = x|Y = y')}$$

이 법칙을 이용해서 무엇을 할 수 있을까?
왜 이 법칙이 중요할까?

베이즈 법칙

Q: 무엇을 할 수 있을까?
A: 베이지언 예측을 할 수 있음
Q: 왜 이게 중요할까?
A: 많은 경우에 P(Y|X)는 모르지만 P(X|Y)는 알고 있을 수 있음

의사의 진단

- 40세의 여성이 유방암 진단을 위해 촬영(mammogram) 검사를 받았다.
- 암에 걸린 것을 y, 유방촬영 결과가 양성인지 음성인지를 x로 나타내보자
- y = 1 암, y = 0 암이 아님
- x = 1 촬영 결과 양성, x = 0 음성
- 암에 걸린 경우 촬영에서 양성이 나올 확률은 0.8
- 암이 아닌 경우지만 양성으로 나올 확률(False positive)은 0.1
- 검사와 상관 없이 40세 여성이 유방암에 걸릴 확률은 0.004
- 검사 결과가 양성이라면 유방암에 걸렸을 확률(posterior probability)은 얼마일까?

의사의 진단

- 40세의 여성이 유방암 진단을 위해 촬영(mammogram) 검사를 받았다.
- 암에 걸린 것을 y, 유방촬영 결과가 양성인지 음성인지를 x로 나타내보자
- y = 1 암, y = 0 암이 아님
- x = 1 촬영 결과 양성, x = 0 음성
- 암에 걸린 경우 촬영에서 양성이 나올 확률은 0.8
- 암이 아닌 경우지만 양성으로 나올 확률(False positive)은 0.1
- 검사와 상관 없이 40세 여성이 유방암에 걸릴 확률은 0.004
- 검사 결과가 양성이라면 유방암에 걸렸을 확률(posterior probability)은 얼마일까?

의사의 진단

- 암에 걸린 경우 촬영에서 양성이 나올 확률은 0.8 (likelihood)

$$p(x = 1|y = 1) = 0.8$$

- 암이 아닌 경우지만 양성으로 나올 확률(False positive)은 0.1

$$p(x = 1|y = 0) = 0.1$$

- 검사와 상관 없이 40세 여성이 유방암에 걸릴 확률은 0.004 (prior)

$$p(y = 1) = 0.004$$

- 검사 결과가 양성이라면 유방암에 걸렸을 확률(posterior probability)은 얼마일까?

Medical diagnosis

$$p(x = 1|y = 1) = 0.8$$
$$p(x = 1|y = 0) = 0.1$$
$$p(y = 1) = 0.004$$

- 검사 결과가 양성이라면 유방암에 걸렸을 확률(posterior probability)은 얼마일까?

$$p(y = 1|x = 1)$$

Medical diagnosis

$$p(x = 1|y = 1) = 0.8$$
$$p(x = 1|y = 0) = 0.1$$
$$p(y = 1) = 0.004$$

- 검사 결과가 양성이라면 유방암에 걸렸을 확률(posterior probability)은 얼마일까?

$$p(y = 1|x = 1) = \frac{p(y = 1)p(x = 1|y = 1)}{\sum_{y'} p(y = y')p(x = 1|y = y')}$$

Medical diagnosis

$$p(x = 1|y = 1) = 0.8$$
$$p(x = 1|y = 0) = 0.1$$
$$p(y = 1) = 0.004$$

- 검사 결과가 양성이라면 유방암에 걸렸을 확률(posterior probability)은 얼마일까?

$$p(y = 1|x = 1) = \frac{p(y = 1)p(x = 1|y = 1)}{\sum_{y'} p(y = y')p(x = 1|y = y')}$$

$$= \frac{p(y = 1)p(x = 1|y = 1)}{p(y = 0)p(x = 1|y = 0) + p(y = 1)p(x = 1|y = 1)}$$

Medical diagnosis

$$p(x = 1|y = 1) = 0.8$$
$$p(x = 1|y = 0) = 0.1$$
$$p(y = 1) = 0.004$$

- 검사 결과가 양성이라면 유방암에 걸렸을 확률(posterior probability)은 얼마일까?

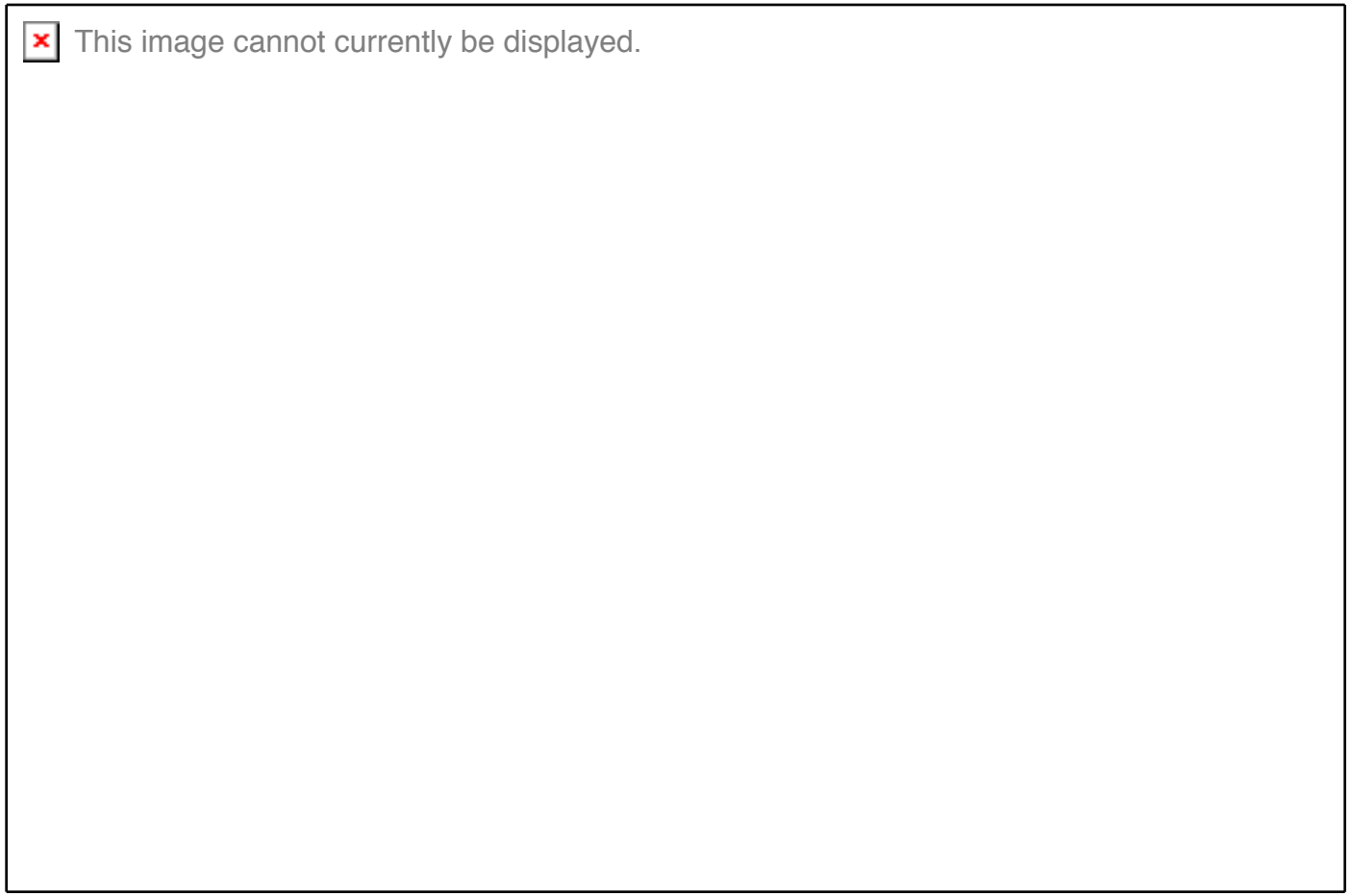
$$= \frac{p(y = 1)p(x = 1|y = 1)}{p(y = 0)p(x = 1|y = 0) + p(y = 1)p(x = 1|y = 1)}$$

$$= \frac{0.004 * 0.8}{0.996 * 0.1 + 0.004 * 0.8} = 0.031$$

베이지언 확률

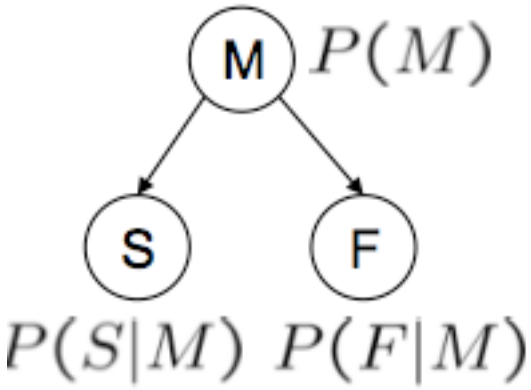
- 의사의 암 진단 문제는 prior probability의 중요성을 강조한다 (암의 회소성이 prior로 들어간다는 것)
- 몬티홀 문제는 likelihood의 중요성과 그 likelihood 확률이 어떻게 문제에서 (호스트의 행동에서) 생성되는지를 보여준다

결혼한 남자일까?



나이브 베이즈 분류법 (Naïve Bayes Classifier)

- 두개의 간단한 binary 특징
  (M: 아들이 있음, F: 패션피플)



나이브 베이즈 분류법 (Naïve Bayes Classifier)

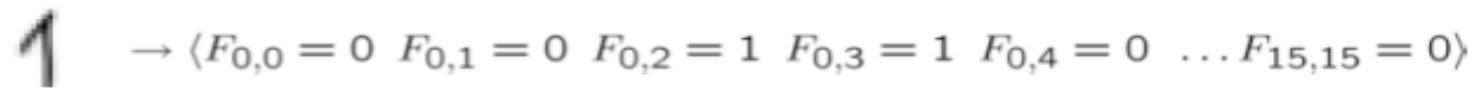
숫자 필기 인식

- 인풋: 픽셀 그리드

- 아웃풋 : 0-9의 숫자

나이브베이즈 분류

- 간단한 방법 :
  - 그리드의 포지션 <I,j>마다 하나의 feature Fij
  - 각 feature의 값은 on / off, 이미지에서 그 포지션의 밝기가 0.5 이상인지에 따라
  - 각 인풋 데이타가 하나의 feature vector로 표현될 수 있음



$$\rightarrow \langle F_{0,0}=0 \;\; F_{0,1}=0 \;\; F_{0,2}=1 \;\; F_{0,3}=1 \;\; F_{0,4}=0 \;\; \ldots F_{15,15}=0 \rangle$$

- 나이브 베이즈 모형:

$$P(Y|F_{0,0} \ldots F_{15,15}) \propto P(Y) \prod_{i,j} P(F_{i,j}|Y)$$

- 여기에서 학습할 것은 무엇일까?

조건부 확률 테이블 (Conditional Probability Table)

조건부 확률 테이블 (Conditional Probability Table)

P(Y) table의 값들이 다 똑같지 않을 수가 있을까?

조건부 확률 테이블 (Conditional Probability Table)

P(F|Y) table의 값들은 어떻게 구할까 ?

General Naïve Bayes

- A general *naive Bayes model:*

보편적인 나이브베이즈 모형

- 일반 나이브 베이즈 모형*:*

- *각 특징(Feature)의 확률이 어떻게 되는지는 Y(클라스)에만 의존한다 (다른 특징들과 별개다)*
- *Parameter의 갯수는 n의 선형 관계이다*

Inference for Naïve Bayes

Goal: compute posterior over classes

나이브 베이즈 모형을 통한 분류

- 목표: 분류될 수 있는 클라스마다 사후확률(posterior) 계산

$$P(Y|F) \propto P(Y) \prod_i P(F_i|Y)$$

나이브 베이즈 모형

- 나이브베이즈 모형을 사용하기 위해 알아야 하는 것은?
- 분류
    - P(Y) and the $P(F_i|Y)$ tables
    - 위의 확률 값들을 사용해 사후 확률 계산 $P(Y|F_1...F_n)$
- 조건부 확률 테이블 구하기
    - P(Y) 각 클라스에 대한 사전 확률
        - $P(F_i|Y)$ 각 특정(feature)에 대한 조건부 확률 (evidence variable)
        - 이 확률값들을 모델의 *parameters* 라고 부르고 *theta 로 표기*
        - 지금까지는 이 값들이 주어졌다고 생각했지만
        - …학습 데이타에서 오는 값들이다

Example: CPTs

Parameter Estimation

- 랜덤 변수의 분포: X or X | Y
- *Empirically:* 학습 data 사용
  - 변수 x의 가능한 값에 대해, 실제로 데이타에서 얼마나 나오는지를 봄 (empirical rate):

  - 이 estimation 방법이 데이타를 가장 잘 나타낸다고 하여 likelihood of the data

$$L(x, \theta) = \prod_i P_\theta(x_i)$$

- *사람 (expert)에게 물어보는 방법 도 있음*

이메일 스팸 분류법

- Naïve Bayes spam filter
- Data:
  - 스팸인지 아닌지 라벨이 붙어있는 이메일 데이타
  - Note: 누군가는 메일을 하나씩 보면서 스팸인지 라벨을 붙여야 함
  - Split into training, held- out, test sets
- Classifiers
  - training set 에 대해 학습, held-out에 대해 튜닝, test set에 대해 정확도 측정

텍스트 분류를 위한 나이브 베이즈

- Bag-of-Words 나이브베이즈 :
  - 스팸인지 아닌지에 대한 분류 (spam vs. ham)
  - 증거 특징(evidence feature), 즉 단어들은 independent하다는 가정

- Generative model

- Bag-of-words
  - 보통 하나의 특징이 하나의 CPT를 갖게 된다 P(F|Y)
  - bag-of-words model 에서는
    - 단어의 position에 대해 따로 보지 않고
    - 단어가 어디에 있던지 같으 단어면 같은 CPT를 갖게 된다 P(W|C)
    - 왜 이런 가정을 할까?

스팸메일 분류

$$P(C, W_1 \dots W_n) = P(C) \prod_i P(W_i|C)$$

- Model:
- 여기에서 parameter 들은?

학습 데이타의 문제

스팸 메일 분류

- 실제 확률값이 아니라 확률값들의 상대적인 ratio가 사후 확률에 차이를 준다

학습 데이타에 대한 문제

- 상대적 확률값을 나타내는 parameter들은 학습 데이타에 overfit하는 문제가 될 수 있다!
  - pixel (15,15)이 "on"이고 숫자 3을 나타내는 인풋이 학습 데이터에 없다고 해서, 테스트 데이타에도 한번도 안 나온다고 단정짓기 어렵다
  - 이메일에 "minute" 이라는 단어가 들어간다고 100% spam ?
  - 이메일에 "seriously"라는 단어가 들어간다고 100% ham ?
  - 학습 데이타에 한번도 나오지 않는 단어는?
  - 일반적으로 보지 못한 feature 값에 대해 0.0의 확률을 주면 안 됨
  - bag-of-words 모델(포지션을 고려하지 않음)을 쓰면 조금은 더 일반적인 결과를 낼 수 있지만, 그걸로 충분하지 않을 수 있음
  - To generalize better: we need to smooth or regularize the estimates

스무딩 (smoothing)

- Problems with maximum likelihood estimates:
  - 동전을 한번 던져서 앞면이 나왔다면 P(heads)?
  - 동전을 10번 던져서 앞면이 8번 나왔다면?
  - 동전을 1천만번 던져서 8백만번 앞면이 나왔다면 ?
- 기본 아이디어:
  - 파라미터에 대한 기본적인 기대치 (prior expectation)
  - 증거(evidence)가 별로 없을때는 기대치에 기울게 됨 증거(evidence)거가 많을때는 데이타에 의존하는게 좋음

라플라스 스무딩 (+1)

- Laplace's estimate:
- 모든 값에 대해 한번 더 봤다고 생각함 (+1)

$$P_{LAP}(x) = \frac{c(x)+1}{\sum_x [c(x)+1]} \qquad P_{ML}(X) =$$

$$= \frac{c(x)+1}{N+|X|} \qquad P_{LAP}(X) =$$

라플라스 스무딩 (더 일반적인 방법)

- Laplace's estimate (extended):
  - 모든 아웃컴에 대해 k번 더 봤다고 생각함

$$P_{LAP,k}(x) = \frac{c(x) + k}{N + k|X|}$$

  - Laplace with k = 0는?
  - k 에 대해 prior의 strength 라고 할수있음
- Laplace for conditionals:
  - 각 condition에 대해 적용:
  - Can be derived by dividing

$$P_{LAP,k}(x|y) = \frac{c(x,y) + k}{c(y) + k|X|}$$

Estimation: Linear Interpolation

- 실제로 conditional P(X|Y)에 대해 laplace smoothing을 적용하면 좋지 않음 (|X|, 특징의 갯수, 혹은 |Y| 클라스의 갯수가 매우 클 경우)
- 다른 옵션 : linear interpolation

$$P_{LIN}(x|y) = \alpha \hat{P}(x|y) + (1.0 - \alpha)\hat{P}(x)$$

알파가 0일때? 1일때?

Tuning on Held-Out Data

- Now we've got two kinds of unknowns
  - Parameters: the probabilities P(Y|X), P(Y)
  - Hyperparameters, like the amount of smoothing to do: k, alpha
- Where to learn?
  - Learn parameters from training data
  - Must tune hyperparameters on different data
    - Why?
  - For each value of the hyperparameters, train and test on the held-out data
  - Choose the best value and do a final test on the test data

Summary

- 조건부 확률과 베이즈 규칙을 이용한 예측
- 나이브베이즈 분류법은 클라스가 주어진다면 각 특징은 conditionally independent 하다고 봄
- 학습 데이타를 갖고 나이브 베이즈 모형을 학습시켜서 분류에 쓸 수 있음
- 실제로 나이브베이즈 모형을 쓸때는 스무딩이 중요함