워크북

교과목명 : 머신 러닝

차시명: 13차시

◈ 담당교수: **장 필 훈**

⊙ 세부목차

- 트리기반 모델조합
  - decision tree
  - 선형회귀의 혼합
- 확률분포
  - 베르누이분포
  - 이항분포
  - 베타분포

학습에 앞서

■ 학습개요

모델조합의 두번째 시간으로 트리를 배운다. 트리로 회귀와 분류 둘 다 가능한데, 구체적으로 어떤식으로 하는 것인지 살펴본다. 트리는 가장 직관적인 모델중의 하나로, 여러 단점에도 불구하고 많이 쓰이는데, 어떤 장점과 단점이 있는지를 자세히 알아본다. 조합의 마지막으로 선형회귀모델의 조합, 로지스틱 및 기타 회귀의 조합을 간단하게 살펴보고, 앞시간에 이미 살펴본것과 같으므로 복습하는 의미로 다시 본다.

모델조합을 마치면 남은 두시간 남짓한 수업은 확률분포에 대해 배운다. 확률분포는 모델링의 기본이고, 여러가지가 연구되어 있으므로 가장 기초적이고 많이 쓰는 분포 위주로 학습한다. 확률분포를 학습하면서, 분포가 하나 주어지면 어떤식으로 해당 분포를 탐구하는지, 반대로 분포를 디자인하려면 어떤 조건들을 만족해야 하고 어떤 성질을 가지면 좋은지, 알려진 분포들은 서로 어떤 관계를 가지는지 자연스럽게 익히는데 중점을 둔다.

■ 학습목표

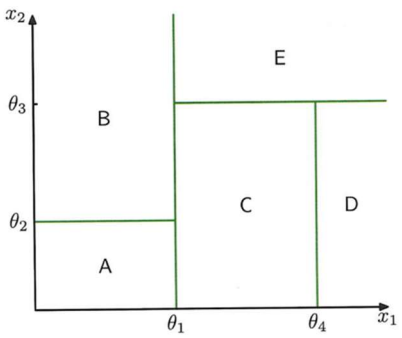
1	트리모델의 특성과 장단점, 구체적인 적용예시(예-결정트리)에 관해 배운다.
2	베르누이 분포의 기댓값과 분산의 계산
3	이항분포의 기댓값과 분산의 계산
4	베타분포의 기댓값과 분산의 계산
5	베타분포와 베르누이분포의 관계, 켤레분포의 정의.

■ 주요용어

용어	해설
결정트리	의사결정트리라고도 하는 지도학습 모델의 하나. 노드와 엣지가 트리형태인 그래프이고, 각 노드마다 T/F로 결정되는 질문을 하나씩 할당할 수 있다. 데이터포인트마다 최종 leaf까지 test가 진행되고 마지막에는 해당 leaf가 가지고 있는 값이나 클래스를 출력으로 내놓게 된다.
베르누이 분포	확률변수 X가 0,1 두가지 값만 가질 때, $P(X=1)=p, 0 \leq p \leq 1$을 만족하는 확률변수X가 따르는 확률분포. 이항분포의 특수한 사례.
이항분포	베르누이 분포에서 본 경우(결과가 0,1인 것)를 반복적으로 실행했을 때의 이산확률분포. 시행이 단 한번일 때가 베르누이 분포이고, X=1인 확률p로 여러번 실행되었을 때 1이 나올 횟수가 따르는 분포가 이항분포이다.
베타분포	이항분포나 베르누이분포처럼 단순하게 정의할 수 없고, 두 매개변수(α, β)에 따라 수많은 분포를 나타낼 수 있게 추상화한 '분포들의 가족'으로 볼 수 있다. 식은 강의록 참고.
감마분포	베타분포와 같이 여러 분포를 대표할 수 있는 추상적인 분포. 매개변수는 두개이고 양의 실수만을 가진다. 식은 강의록 참고.

	학습하기

모델조합중의 하나로 트리기반 모델의 간략한 개념을 살펴보겠습니다. 먼저 2차원 공간을 트리를 사용해 나누었다고 가정해보겠습니다. 구분 기준은, 무작위적으로 입력차원중 한개차원을 골라서 일정한 기준값으로 구분했다고 가정해보겠습니다.



[Bishop. Fig.14.5]

트리로 분할된 각 구역 안에서 또 다른 각각의 개별모델을 사용해 예측할 수 있습니다. 회귀라면 단순 상수를 출력으로 주어야 할 것이고(한 구역 안에서 서로 다른 값을 줄 수 없습니다. 그러려면 구역을 나누어야 합니다. 또한 일정한 함수와 그에 따른 출력값을 줄수도 없습니다. 그렇게 되면 트리기반 모델이 아니게 됩니다. 물론 그렇게 디자인하고, 모델혼합이라고 이름붙이면 아무 문제 없습니다. 하기 나름입니다), 분류라면 구역 내 모든 포인트가 하나의 클래스에 해당해야 할 것입니다.

트리기반 모델의 장점은 해석이 쉽다는 것입니다. 해석이 쉬우면 모델을 수정하기도 쉽습니다.

트리를 디자인할 때는 구조, 분할기준, 할당값등 고려해야할 것이 몇가지 있지만, 얼마나 복잡하게 분할할 것인가가 주로 문제가 됩니다. 극단적으로, 오버피팅하려고 마음먹고 모든 데이터포인트가 하나의 구역에 대응하게 트리의 가지를 늘려나갈수도 있습니다. 이렇게 하면 오류는 적어지겠지만 복잡도가 지나치게 증가합니다. 반대의 경우는 오류를 늘릴 것입니다. 이 둘의 트레이드오프라고 생각하면 됩니다.

회귀문제를 예로 들자면, 하나의 구역에 몇개의 데이터포인트가 최종적으로 들어갔을 때 해당 구역이 주어야 하는 출력값은 평균이어야 제곱오차를 최소화할 것입니다. 그러면 제곱오차와 복잡도를 임의의 계수로 연결하고 가지치기 기준값을 삼습니다. 이 기준값을 최소화 하는 부분에서 가지치기를 멈추면 됩니다.

분류문제의 경우는 회귀문제와 거의 동일하나, 성능척도가 다릅니다. 보통은 크로스 엔트로피나 지니 인덱스를 사용합니다. 이 둘에 관해서 더 많이 궁금하시면 강의를 참고하셔도 좋고 더 자세한 자료가 웹에 많으니 참고바랍니다.

<결정트리>

아마 트리기반 모델이라고 하면 가장 많은 사람들이 결정트리를 생각할 것입니다. 간단히 이야기하면, 예/아니오로 나뉘는 수많은 가지로 이루어진 트리가 결정트리입니다. 물론 입력이 연속적 값의 경우는 예/아니오가 아니라 특정값이 기준이 되겠습니다.

선형회귀의 혼합과 로지스틱회귀등의 혼합은 중요도가 떨어져서 자세히 설명하지 않습니다. 강의시간에 다루었으니 참고하시면 됩니다.

<확률분포>

지금까지는 머신러닝분야의 고전적인 이론들을 습득했습니다. 이 배경에 베이지안적 접근이 많이 사용되었고, 그렇지 않더라도 확률에 관한 기본적인 개념이 많이 필요했습니다. 우리는 대부분 가우시안만 사용했기 때문에 머신러닝을 배우기 위해 꼭 여러 확률분포에 대해 자세한 지식이 필요하지는 않습니다. 지금부터는 확률에 대한 기본을 배운다는 생각으로 살펴보시면 되고, 중간에 나오는 복잡한 식은

스킵하셔도 좋습니다.

여러 확률분포에 관해 학습하는것은 대략의 루틴(평균, 분산계산, 파라미터변환에 따른 분포의 변화 등)이 있습니다. 그래서 참고할 자료도 구하기가 쉽습니다. 더 좋은 통계 강의들이 많이 있으니, 다른 스타일의 강의를 원하시는 분들은 다른강의를 참고하셔도 좋겠습니다.

<베르누이분포>

베르누이분포는 확률변수가 두개의 값(0,1)만 가집니다. 그리고 1일확률을 μ라고 둡니다.

그러면 0일 확률은 당연히 $1-\mu$가 되겠지요.

이런 시행을 여러번(x번) 한다고 생각하면, 분포는 다음과 같아집니다.(여러 사건의 동시발생확률을 계산하는것과 같습니다)

$$\mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$$

평균: $\mathbb{E}[x] = p_{x=1}\cdot 1 + p_{x=0}\cdot 0 = \mu\cdot 1 + (1-\mu)\cdot 0 = \mu$

분산: $\mathbb{E}[x^2] = p_{x=1}\cdot 1^2 + p_{x=0}\cdot 0^2 = \mu = \mathbb{E}[x]$

$$\mathrm{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \mu - \mu^2 = \mu(1-\mu)$$

가능도함수는 위 확률분포의 곱과 같습니다.(확률변수 x는 iid, 평균이μ인 데이터집합 D에서 추출)

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

로그가능도 함수: $$\sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln\mu + (1-x_n)\ln(1-\mu)\}$$

위 함수를μ에 관해 미분하고 0으로 둡니다(MLE)

최대가능도 추정값: $$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

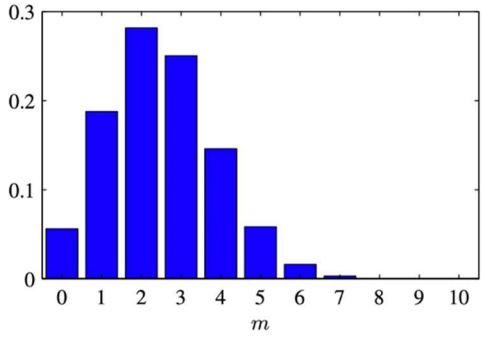
N이 작을 때 왜곡이 쉽다는것이 단점입니다. 예를들어 동전을 세번 던졌을 때 모두 앞면이면, μ_{ML} = 1이 됩니다. 앞으로 계속 앞면이라는 뜻이죠. 이럴 때 베이지안 관점이 필요합니다. μ에 관해서도 사전분포를 가정하는 것입니다. (이에 관해서는 뒤에 다룹니다)

<이항분포>

이항분포는 binomial distribution으로, 확률변수 x가 0,1로 주어질 때 N번 시행후 1이 몇개나 되는지에 대한 분포입니다. 확률(조합)로 단순히 구할수 있고, 다음과 같이 적습니다.

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}, \quad \binom{N}{m} \equiv \frac{N!}{(N-m)!\,m!}$$

N=10, μ=0.25를 넣고 실제로 계산해보면 분포가 다음과 같습니다.



[Bishop. Fig.2.1]

평균은 다음과 같습니다.

$$\mathbb{E}[m] = \mathbb{E}[x_1 + x_2 + \cdots + x_N] = \mu + \mu + \cdots + \mu = N\mu$$

분산을 구하기 위해서 다음을 이용합니다.

$$\mathrm{Var}\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} \mathrm{Var}(X_j)$$ (독립시행을 가정했을 때 왼쪽관계가 성립. 증명은 강의 참조.)

따라서 분산은 다음과 같습니다.

$$\mathrm{Var}[m] = \mathrm{Var}[x_1 + x_2 + \cdots + x_N] = \mathrm{Var}\left(\sum_{i}^{N} x_i\right) = \sum_{i}^{N} \mathrm{Var}(x_i)$$

Var(x)= $\mu(1-\mu)$이므로,

$$\mathrm{Var}[m] = \sum_{i}^{N} \mu(1-\mu) \; = N\mu(1-\mu)$$

<베타분포>

매개변수μ에 대한 사전분포 p(μ)를 가정하고, 베타변수는 다음과 같이 정의됩니다.

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(\mathrm{b})}\mu^{a-1}(1-\mu)^{b-1} = \frac{1}{\mathrm{B}(a,b)}\mu^{a-1}(1-\mu)^{b-1}$$

B(a,b)는 베타함수로서 정규화 역할을 합니다. 베타함수는 다음과 같습니다.

$$\mathrm{B}(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$$

a, b는 hyperparameter입니다. 임의로 정할 수 있습니다.

그리고 감마함수는 다음과 같이 정의됩니다.

$$\Gamma(x) \equiv \int_0^\infty u^{x-1}e^{-u}du \; = \lim_{m\to\infty} \frac{m^x m!}{x(x+1)(x+2)\cdots(x+m)}$$

두번째 형식은 오일러 폼이라고 하는데, 좀 더 익숙하게 다가올 것 같아 쓰긴 했지만, 그렇지 않다면 무시하시면 됩니다.

기댓값을 구해보겠습니다.

$$E(x) = \int_0^1 x \cdot \frac{x^{a-1}(1-x)^{b-1}}{\mathrm{B}(a,b)} dx = \frac{1}{\mathrm{B}(a,b)} \int_0^1 x^a (1-x)^{b-1} dx$$
$$= \frac{\mathrm{B}(a+1,b)}{\mathrm{B}(a,b)} = \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

감마함수는 오른쪽과 같이 전개해보면, $\int_0^\infty u^x e^{-u} du = [-u^x e^{-u}]_0^\infty + x \int_0^\infty u^{x-1} e^{-u} du = x\Gamma(x)$

$\Gamma(x+1) = x\Gamma(x)$로 쓸 수 있습니다.

따라서 위 기댓값을 계속 전개하면 다음과 같습니다.

$$\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \frac{a}{a+b}$$

분산은 다음과 같습니다.

$$\mathrm{Var}(x) = \frac{1}{\mathrm{B}(a,b)} \int_0^1 x^{a+1}(1-x)^{b-1} dx - \frac{a^2}{(a+b)^2}$$

기댓값을 구할때와 마찬가지로 베타함수를 감마함수로 바꾸고, 서로 cancel out시킵니다.

$$= \frac{\mathrm{B}(a+2,b)}{\mathrm{B}(a,b)} - \frac{a^2}{(a+b)^2} = \frac{ab}{(a+b)^2(a+b+1)}$$

위에서 우리는 이항분포에서 평균의 사전분포를 정의할 수 있다고 했습니다. 그 사전분포를 베타함수로 정의하고 가능도 함수를 구해보겠습니다.

$$P(m|\mu) = \binom{M}{m} \mu^m (1-\mu)^{M-m}$$
$$P(\mu) = \frac{\mu^{a-1}(1-\mu)^{b-1}}{\mathrm{B}(a,b)}$$
$$P(\mu|m) \propto P(m|\mu)P(\mu) = \binom{M}{m} \frac{\mu^{m+a-1}(1-\mu)^{M-m+b-1}}{\mathrm{B}(a,b)}$$

사후분포의 모양이 사전분포와 같음을 볼 수 있습니다. 즉, 사후분포도 베타함수가 되었습니다.

새로 구한 식을 사전분포와 비교해보면, a가 m만큼, b가 M-m만큼 증가한 것으로 이해할 수 있습니다. 추가관측이 이루어질때마다 계속 분포를 업데이트(사후분포가 다시 사전분포가 되는 과정이 반복됨)할 수 있다는 뜻입니다.

	연습문제

1. 트리기반 모델을 회귀로 사용하려면, 각 구역 내에서는 상수출력을 줄 수 밖에 없다.
   a. O
   b. 함수를 출력으로 주는 형태가 아닌한, 트리모델은 불연속적인 출력을 특징으로 하므로 각 구역 내에서는 한가지 값 밖에 줄 수 없다. 따라서 상수출력이 된다.
2. 트리기반모델에서 특정 깊이 이상으로 분할했을 때 오류를 감소시킬 수 없는 지점이 있다면, 그 지점이 최소오류를 달성해내는 지점이므로 해당지점에서 분할을 멈추면 된다.
   a. X

b. 어떤 단계에서 오류를 감소시킬 수 없었더라도 더 진행하면 오류가 감소하는 상황이 가능하다. 따라서 어떤 지점에서 분할을 멈추는 것이 최선인지 일반적 상황에서 계산해보기 전에 정확히 예측하는것은 불가능하다고 알려져 있다.

3. 관찰집합이 주어졌을 때, 최적의 유일한 확률분포를 찾는 것은 수학적으로 불가능하다.
   a. O
   b. 가능한 모델은 무한하기 때문에, 그중에 어떤 것이 '답'인지는 선택의 문제이다. 따라서, 해가 존재하더라도 '유일한'해가 아니기 때문에 지문은 옳다.
4. 베르누이분포의 기댓값은 x=1일 확률μ와 동일하고, 분산은 $\mu(1-\mu)$이다.
   a. O
   b. 계산하면 나옴. 강의록 참고.
5. 베르누이분포를 따르는 확률변수의 관측값을 바탕으로 가장 가능성 높은 μ를 구해보면, 관측값 중 1의 비율과 일치한다.
   a. O
   b. 강의록 참고. $\mu_{ML} = \frac{m}{N}$이다.
6. 데이터의 숫자를 N, 1이 나올 확률을 μ라고 하면, 이항분포의 평균은 $N\mu$이고, 분산은 $N\mu(1-\mu)$이다.
   a. O
   b. 강의록 참고.
7. 가능도함수가 이항분포를 따르고 사전분포가 베타함수를 따르면 사후분포도 베타분포를 따른다.
   a. O
   b. 강의록 참고. 식으로 증명이 가능하다.

정리하기

1. 트리기반 모델은 구역을 불연속적으로 구분하게 된다.
   a. 회귀의 경우 각 구역마다 대표값이 정해지기 때문에 공간 전체에 대해 보면 출력값들이 불연속이게 된다.
   b. 분류의 경우 각 구역이 특정 클래스로 배정되게 되므로 같은 구역에 속하는 모든 데이터 포인트는 모두 같은 클래스로 분류된다.
2. 트리모델을 만들 때는 다음 두가지가 문제된다.
   a. 어떤 분류기준을 삼는것이 최적인지
      i. 구역에 속하는 데이터포인트에 대해 가능한 모든 경우 중 가장 분류나 회귀 오류를 최소화 하는 분할을 선택하는 것이 가장 보편적이다.

ii. 따라서 계산량이 많다.

b. 언제 분할을 멈출 것인지.

i. 어떤 분할도 오류를 감소시키지 않는 단계가 나왔을 때 무조건 중단하는 것이 가장 최적의 선택은 아니다. 거기서 더 진행했을 때 오류를 줄이는 경우가 있다.

ii. 너무 분할을 많이 하게 되면 모델의 복잡도가 쓸모없이 증가하고 오버피팅의 위험이 생기므로 그 둘의 균형점을 잘 찾아서 나눈다. 이 경우에도 여러가지 시도해보고 결정하는 때가 많다.

3. 데이터 포인트만 보고 그 데이터 포인트가 어떤 분포로부터 추출되었는지 알아내는 것이 확률분포문제

a. 하지만 실제 문제가 적용되는 환경에서 데이터가 그렇게 수학적으로 말끔한 모양의 분포로부터 추출되는 경우는 없다.

b. 따라서 가장 비슷하고, 해석하기 쉬운 모양으로 확률분포를 추정해 내는 것이 과제가 된다.

c. 데이터포인트만 보고 모델을 추정해 내는 것은 해답이 이론적으로 무한하므로, 결국 모델선택의 문제로 귀결한다.

4. 베르누이 분포, 이항분포, 베타분포, 감마분포의 기댓값(평균)과 분산을 계산하는 방법은 강의록 참조.

a. 어떤 확률분포든 기본적으로 계산하게 되는 통계량

5. 사전분포와 가능도함수를 알면 사후분포를 얻어낼 수 있는데, 우리가 공부하는 여러 분포가 이 관계에 얽혀있다. 예를들어 가능도함수가 이항분포, 사전분포가 베타함수면 사후분포도 베타함수를 따른다. 정확한 계산은 강의록을 참고할것.

참고하기

Bishop, C. M. "Bishop-Pattern Recognition and Machine Learning-Springer 2006." Antimicrob. Agents Chemother (2014): 03728-14.

다음 차시 예고

- 다항변수
- 디리클레분포
- 가우시안분포
- 감마분포