1.5 Decison Theory

To make optimal decisions in situation involving uncertainy

input X 와 output t 에 대한 joint dist 를 찾아낸다면 온전히 변수간의 관계를 알아내는 것이므로 불확실성을 완전히 통제할 수 있을 것이다. 이것을 inference라고 한다.

그러나 결정이론에서 관심있는 바는 specific prediction이므로 joint dist 로부터 확률을 얻어내어 optimal decision을 얻으려고 한다.

즉, 결정이론과 확률이론이 여기에서 연결이 된다. 확률을 가지고 결정을 하고 싶은데 이 때 관심있는 확률은 사후확률

$$ p(t | X) = p(C_k|X) = \frac {p(X|C_k)p(C_k)}{p(X)} $$

이며 오류를 최소화하기 위해서는 사후확률이 가장 큰 class를 선택하는 것이 바람직할 것이라 예상된다.

(사전확률 : X-ray 관찰 이전에 어떤 사람이 암을 가지고 있을 확률)
(사후확률 : X-ray 관찰 이후에 어떤 사람이 암을 가지고 있을 확률)

1.5.1 Minimizing the misclassification rate

relative relative

오류의 확률을 줄이는 것이 목적이라면 $ p(X , C_1) > p(X, C_2) $ 일 때 C_1 을 선택하는 것이 바람직하다. Bayes rule을 적용할 때 공통의 $p(X)$에 대해 사후확률 $(C_1|X) > p(C_2|X)$ 를 선택하는 것과 동일하다. 또한 이와 동치의 사건은

relative

의 확률을 극대화하는 것이다.

즉 두가지 경우 모두 사후확률이 가장 큰 class를 선택할 때 오류를 줄일 수 있음을 보여준다.

decision region , decision boundary , minimize error 를 한꺼번에 나타내는 그림이 다음과 같다. relative

$\hat x$ 가 decision boundary 를 나타내고 x 축에 나오는 R_1 R_2 가 decision region을, 색깔로 칠해진 부분은 두 가지 error를 나타낸다. 적분을 통해 쉽게 유추가능한데, 붉은 색으로 칠해진 부분은 boundary를 바꿈으로써 줄일 수 있기 때문에 optimal 을 쉽게 찾을 수 있다. 이렇게 구한 optimal이 결국 사후분포를 최대로 하는 영역에 class를 할당하는 것이다.

1.5.2 Minimizing the expected loss

1종의 오류를 줄이는 것이 중요하기 때문에, loss matrix에서 오류마다 가중치를 다르게 주어서 loss function 이나 cost function을 만든다. L_{kj}는 loss matrix 의 원소

$$ E[L] = \sum _ k \sum_j \int_{R_j} L_{kj}p(x,C_k) $$

이를 최소로 하고 싶은데 j 값은 우리가 선택한 영역이므로 j 의 값과 상관없이

$$ \sum _ k L_{kj}p(x,C_k) $$ 를 최소로 하는 decision boundary 를 찾거나 혹은 bayes rule을 사용해서

$$ \sum _ k L_{kj}p(C_k|X) $$ 를 최소로 하는 decision boundary를 결정한다.

1.5.3 The reject option

사후확률을 기준으로 볼 때, θ 보다 높은 값을 값을 가지면 올바르게 분류할 확률이 높은 것이므로 양 끝 부분을 제외한 사이부분이 기각역이 된다. <그림참고>

relative

1.5.4

1. inference stage (generative model)

joint dist 와 사후분포를 통해서 확률을 중심으로 한 모델을 만드는 것.

데이터로 부터 joint dist를 찾아내는 것은 쉽지 않다. 그러나 bayes rule 을 통해서 알아낸 $p(X)$는 이상치 탐지에 도움을 준다.(실제 데이터가 일어날 확률에 대해 알려주기 때문)

2. decision stage (discriminative model)

inferecne 이후에 decision rule을 통해서 class를 할당하는 것.

joint dist 와 베이즈 규칙을 통해 확률을 찾는 것은 계산량도 많고 비효율적이다.

3. inference + decision (discriminant function)

discriminant function을 찾아서 분류를 하는 것. 여기서 표면적으로 볼 때 확률은 거의 사용되지 않는다.

1,2,3 세가지 관점을 모두 대략적으로 보았는데 사전분포를 아는 것은 확실히 효과가 있긴 하다. 그 이유는 아래와 같다.

1)우선 loss matrix를 바꿔서 다양하게 decision boundary 를 바꿀 수 있다.

2)또한 사후분포를 알고 있다면 앞에서 말한대로 misclassification rate 이나 expected loss를 구할 수 있다.

3)unbalanced data 에서는 bayes rule에 의해 이를 고려한 사후분포를 만들어 낼 수 있다.

4)복잡한 문제를 나눠서 생각할 수 있다(naive bayes model - conditional independence model)

특히 4 번의 경우

$$ \begin{align} p(C_k| x_1, x_2) \propto p(x_1,x_2 | C_k)p(C_k) \ \propto p(x_1|C_k)p(x_2|C_k)p(C_k) \ \propto \frac{p(C_k|x_1)p(C_k | x_2)}{p(C_k)} \end{align} $$

(의미 : 두 가지 변수 x1,x2 가 동시에 일어났을 때 특정 class 로 분류할 확률은 각 변수가 독립인 경우에 각 변수 개별로 조건부를 취한 사후분포의 곱과 비례한다.)

1.5.5 Loss function for regression

지금까지는 분류 문제를 가지고 loss function 과 decision theory를 적용시켜 보았다면, 여기서는 regression 문제에서 이를 살펴보자.

$$ E[L] = \int \int L(t,y(x))p(x,t) dx dt $$

에서 이를 최소로하는 y function을 구하기 위해 미분을 취하면, squared loss 인 경우 y(x)를 미분한 값인

$$ 2 \int { y(x)- t }\ p(x,t) dt = 0 $$

를 최소로 하는 값으로 선택하면 된다.

squared loss 이므로 optimal solution은 E[t|x]

loss에 대해 아래와 같은 방식으로 접근해볼 수 있다.

relative relative

여기서 second term 은

variance of the dist of t, averaged over X : intrinsic variability , noise, irreducible minimum value of the loss fuction

앞에서 quadratic loss 만 봤는데 이를 일반화 시킬 때

$$ E[L_q] = \int \int |t,y(x)|^q p(x,t) dx dt $$

를 Minkowski loss라고 하고 cost fuction을 최소로 만드는 y(x)는 q=2 일 때 조건부 기댓값, q=1 일 때 조건부 중위값, q가 0과 1 사이일 때 조건부 최빈값을 가진다.

1.6 Information Theory

희귀한 정보일수록 정보량이 많다! 따라서 희귀성을 판단하기 위한 $p(x)$ 와 정보량을 나타내는 $h(x)$ 도입

두 정보가 독립일 경우 $$ h(x,y) = h(x) + h(y) \ p(x,y) = p(x)p(y) $$

이고 여기서 h와 확률값이 log 에 의해 연결되어 있는 것을 알 수 있다.

$$ h(x) = -log_2(p(x)) $$

만약 평균적인 정보량을 알고싶다면? $$ = - \sum_x p(x)log_2(p(x)) $$

를 통해 구하는데 이를 확률변수 x 에 대한 entropy라고 한다.

위 공식에 의하면 non-uniform dist 일 때 정보량이 더욱 작아진다는 것을 충분히 예상할 수 있다.

1. entropy가 언제 최대가 되는지 수식으로 쉽게 증명이 가능하다. (라그랑지안 사용)

$$ \tilde H = - \sum p(x_i) \ln p(x_i) + \lambda (\sum p(x_i) - 1) $$

relative

2. 또다른 증명 방법으로는 jensen's inequality를 사용하는 방법이 있다.

$$ -logx $$

는 convex function이므로

$$

- \sum p(x) log \frac {1}{p(x)} \leq log(n) $$

이고 log(n)은 x 가 uniform dist 일 때 나오는 값이다.

(※ entropy 와 coding 간의 관계 : entropy 가 더 작을 수록, 즉 정보량이 더 작을 수록 짧은 고딩으로 정보를 전달할 수 있다 / entropy의 양이 coding 길이의 lower bound)

또한 연속형 데이터에 대해서는 정보량을 differential entropy 로 정의할 수 있는데 (형태는 똑같다. 차분 의미가 들어가서 differential 인 듯)

$$ = - \int p(x) \ln (p(x)) dx $$

또한 이산형에 대한 entropy의 최댓값을 구하기 위해서는 추가적으로 1차 , 2차 moment에 대한 조건이 같이 고려되어야 하기 때문에 다음과 같은 식을 최대화한다.

relative

수식을 푼 결과 differential entropy를 최대화 하는 분포는 *가우시안 분포*이다.

이 때 entropy는

$$ = \frac {1}{2} {1 + \ln(2\pi \sigma^2)} $$

이며 분산 값이 증가할수록 entropy 값도 증가하는 것을 알 수 있다. 또한 이산형과는 다르게 음의 값 가질 수 있다.

Conditional Entropy

average additional information $H[y|x] = - \int \int p(y,x) \ln p(y|x) dy dx $

이고 이를 활용할 때

$$ H[x,y] = H[y|x] + H[x] $$

가 성립한다. 즉 x, y 를 설명하는 정보의 양은 x 를 설명하는데 필요한 정보의 양에 조건부 entropy를 더한 값이다.

1.6.1 Relative entropy and mutual information

unknwon distribution \$p(x)\$ * *true*

approximationg distribution \$q(x)\$

relative entropy = Kullback-Leibler divergence = additional amount of information required to specify the value of x as a result of using q(x) = measure of the dissimilarity

\$\$ KL(p||q) = -\int p(x) \ln q(x) dx - (-\int p(x) \ln p(x) dx) \$\$

특징

1. 양수이다
2. 등호는 p = q 일 때 성립한다.

증명은 다음과 같다.

relative

쿨백 라이블러 divergence를 줄이는 방식으로 true dist p(x) 를 q(x|parameter)를 통해 찾을 수 있을 것이다. 모수적 방법으로 접근했을 때, KL(p||q)를 최소화 시키는 q(x|parameter)를 구하는 것은 결국 q(x|parameter)를 maximize 하는 것과 같다. 즉 likelihood function을 최대화 하는 것과 동일한 것이다.

mutual information (변수들이 서로 독립인지 아닌지 여부를 KLD로 판정한다)

relative

식을 통해 볼 때 I(x,y) 는 항상 0 이상의 값을 가지며 등호를 만족하는 것은 결국 x와 y가 서로 독립임을 의미한다.

또한 mutual information은 entropy를 가지고 표현이 가능하다(이는 확률의 법칙에 의한 것으로 그냥 단순 계산이다)

relative

이것의 의미는 기존 x 에 대한 정보(entropy)에서 y 가 주어졌을 때 x 의 정보(conditional entropy) 를 뺀 것이다. 이는 베이지안의 관점과 유사하다고 할 수 있다. 다시 말해, 기존 prior 에서 y 값이 관찰되고 난 후 x를 보는 posterior 를 뺀 차이라 여길 수 있다.