ML IN PROBABILISTIC PERSPECTIVE

Week 1. OT

강경훈

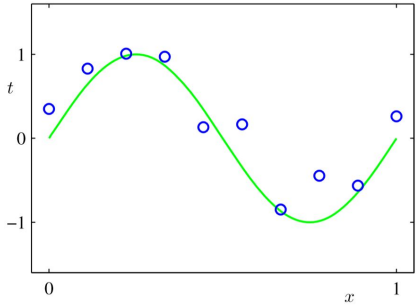
ESC, YONSEI UNIVERSITY

April 1, 2020

Table of Contents

CURVE FITTING EXAMPLE

- 아래의 에시를 통해 머신러닝의 주요 개념을 살펴보자. 우리에게 주어진 데이터는 다음과 같다.



- 데이터 t의 형성 과정은 포괄적으로 말하면 $t = f(x) + \epsilon$로 볼 수 있다. 여기서 $f(x)$는 초록선을, ϵ은 초록선 위주로 생긴 오차로 볼 수 있다.
- 우리의 목적은 주어진 데이터 (x, t)를 통해 최대한 초록선과 가까운 선 $\hat{f}(x)$을 그어, 새로운 데이터 $\hat{x}$가 주어졌을 때 $\hat{t}$를 예측(fitting)하는 것. How?

CURVE FITTING EXAMPLE

- 중요한 것은 관측 데이터에는 항상 노이즈가 껴있다는 것! 관측오차일 수도 있고, x와 상관 없는 변수의 영향일 수도 있다. 데이터가 가지는 패턴 $f(x)$를 잘 포착하되, 노이즈 ϵ는 걸려야 한다.
- 여러 방법이 있겠지만, 일단 이론 없이 직관적으로 생각해보자. $f(x)$를 다항식으로 가정해보자.

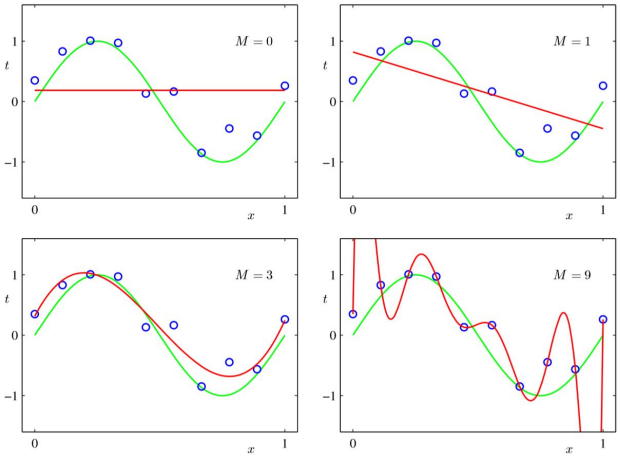
$$f(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M = \sum_{j=0}^{M} w_j x^j = \mathbf{x^T w}$$

  결국 $f(x)$를 정하는 것은 먼저 1) 차수 M를 정하고, 2) 계수 $\mathbf{w}$를 구하는 과정이다.
- M이 주어졌다고 해보자. 모든 데이터에 걸쳐 예측값 $\hat{y}$과 실제값 y의 거리를 error function $E(\mathbf{w})$으로 정의해보자. 그리고 이 거리를 최소화하는 계수 $\mathbf{w}^*$를 구해보자. $E(\mathbf{w})$이 벡터 $\mathbf{w}$에 대한 2차식이므로 미분하면 1차식이니, 해를 간단히 구할 수 있다.

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} \{f(x_n, \mathbf{w}) - t_n\}^2 = \arg\min_{\mathbf{w}} \|\mathbf{Xw} - \mathbf{t}\|^2$$

CURVE FITTING EXAMPLE

- 그렇다면 차수 M을 어떻게 정할까? 일단 한 번 죄다 그려보자.



CURVE FITTING EXAMPLE

- 차수가 너무 높아져도 안 되고, 너무 낮아도 안 된다. 적당한 차수를 어떻게 정할까? 각 차수별로 그래프가 데이터를 얼마나 잘 fit하는지에 대한 척도가 있으면 좋겠다. 정의해보자.

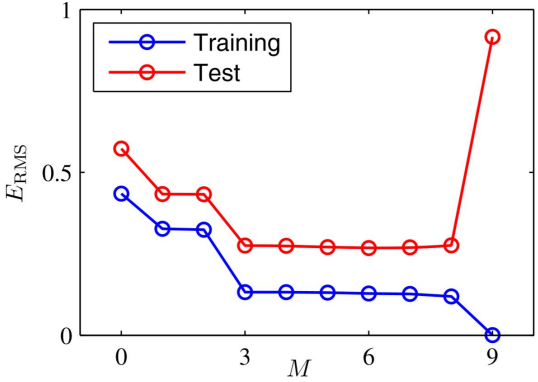
$$Err_{RMS} = \sqrt{Err(\mathbf{w}^*)/N}$$

  Err_{RMS}가 낮을수록 좋은 모델이 아닐까? 제일 낮은 M으로 정하면?
- 그러나 차수를 계속 높이다보면 $Err_{RMS} = 0$, 모든 데이터를 통과하는 선을 얻게된다. 그렇지만 이 경우 다른 데이터에 fitting 해보면 예측이 크게 변동한다. 데이터를 정확히 통과하려고 (에러를 낮추려고) 하다보니 데이터가 가지고 있는 에러도 읽어버렸으니, 다른 데이터를 가져오면 크게 틀리는 것.
- $M = 9$일 때의 계수를 보면 왜 이런 결과가 나온지 이해가 간다. 데이터에 "finely tuned" 하려다보니 차수가 올라갈수록 계수가 지멋대로 커져버린다. (**Overfitted**)

$$w^*_{M=9} = [0.35, 232.37, -5321.83, ..., -1061800.52, ..., 125201.43]$$

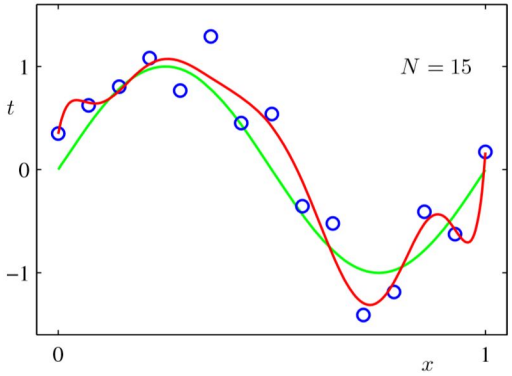
CURVE FITTING EXAMPLE

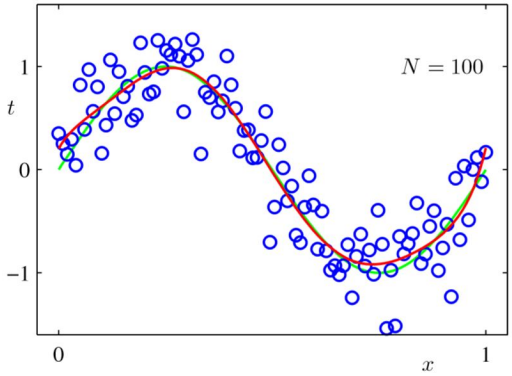
- Train set은 내가 갖고 있는 데이터, Test set은 어디서 또 갖고 온 다른 데이터이다. 과적화된 모델은 Test RMSE가 되려 높아진다.



CURVE FITTING EXAMPLE

- 이런 결과가 나온 이유 중 하나는 데이터 개수 n에 비해 차수 p가 너무 높은 것. 데이터가 많으면 또 이쁘게 잘 나온다.





CURVE FITTING EXAMPLE

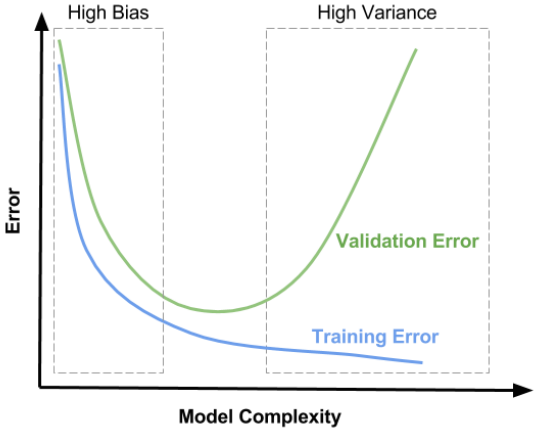
Model Complexity and Bias-Var Tradeoff

- M을 결정한다는 것은 **1) 모델의 복잡도**를 결정하는 것이며, M의 차수를 올릴 때마다 새로운 설명변수가 추가되는 셈이니 **2) 설명변수 feature의 개수**를 결정하는 것으로 볼 수 있다.
- 너무 단순한 모델, 예컨대 그냥 직선 하나 찍 긋는 것은 당연히 실제의 구불구불한 선과 차이가 크다. 이것을 모델의 **Bias**라고 한다.
- 그러나 너무 복잡한 모델, 예컨대 9차 다항식을 그러버리면, 데이터가 달라질 때 모델의 예측값이 크게 널뛰기 한다. 이것을 모델의 **Variance**라고 한다.
- 우리가 관측할 수 있는 Train MSE는 모델이 복잡할수록 항상 줄어들 수 밖에 없다. 그러나 대개의 경우 모델이 복잡해지면 정작 중요한 Test MSE는 어느 정도까지는 줄어들다가 (Bias ↓) 다시 올라간다 (Variance ↑). 이를 **Bias-Variance Tradeoff**라고 한다.
- 때문에 Model Complexity를 결정할 때에는 데이터의 개수를 고려해 $n >> p$가 되도록 함이 바람직하지만, 항상 데이터가 충분하면 얼마나 좋겠나. 때문에 적은 데이터로 최대한 '자연스러운' 결과를 내는 방법들이 중요하다.

CURVE FITTING EXAMPLE

Model Complexity and Bias-Var Tradeoff

- 가운데의 골디락스 존을 찾아라! (출처)



CURVE FITTING EXAMPLE

Test Set이 없다면? Test MSE를 구하기 위한 Resampling

- **주어진 데이터서 반복적으로 "Resample", 왜?**
  ex) 주어진 데이터에 일단 fitting를 하긴 했는데, 이게 다른 데이터를 집어넣으면 얼마나 널뛰기할까가 궁금하다.
  → 여러 데이터로 fitting을 해 회귀계수를 마니마니 구해보자. 이것들을 보면 회귀계수의 sampling variability에 대해 가늠할 수 있다. 근데 난 데이터 하나밖에 없는데?
  → **데이터를 쪼개자!!**
- 하나의 데이터를 여러 개로 쪼개 fitting을 여러 번 한다(추정치를 여러 번 구한다.).
  모델의 성능과 추정치의 분산에 대해 더 많은 것을 알 수 있다!

대표적으로 쓰이는 방법은 **Cross-Validation, Bootstrap**

1. **Cross-Validation** : 데이터를 train/test 셋으로 여러 번 나눠 여러 번 test MSE를 구해, 1) 이 모델이 잘 맞는가 2) 어느 정도로 빡세게 fitting해야 하는가를 알아보자.
2. **Bootstrap** : 데이터에서 새로 랜덤으로 추출한 미니 데이터로 추정치를 잔뜩 구해, 데이터에 따라 추정치가 얼마나 널뛰는가 보자.

Table of Contents

CURVE FITTING IN PROBABILISTIC PERSPECTIVE

- 머신러닝은 기본적으로 "uncertainty"에 대한 학문이다.
  1) 한정된 데이터로 일반화해야하며, 2) 그 데이터에 심지어 에러가 묻어있기 때문이다.
  즉 $f(x)$가 뭔지도 모르고, 데이터에서 $f(x)$와 ϵ을 구분하기도 힘들다.
- Probability theory는 이러한 불확실성을 정량적으로 분석할 수 있는 틀을 제공하며, Decision Theory는 이러한 토대를 바탕으로 결정을 내릴 수 있는 기준을 준다
- **머신러닝을 확률적으로 이해한다는 것은 데이터를 어떤 분포를 가진 확률변수로 보는 것.**
  가우시안 에러를 가정하면 결국 $f(x)$는 x가 주어졌을 때 t의 조건부평균으로 볼 수 있으며, 데이터 하나의 조건부분포는 다음과 같다.

$$\textbf{Sampling Density} \quad p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|\mathbf{x}^T\mathbf{w}, \beta^{-1})$$

즉 하나의 t 데이터가 가지고 있는 모든 불확실성을 하나의 pdf로 가정한 것. 데이터가 어떻게 나왔냐를 설명해주는 sampling density라고도 한다.

CURVE FITTING IN PROBABILISTIC PERSPECTIVE

- 데이터를 전체 N개의 iid sample의 집합으로 가정한다면, 전체 데이터 셋의 분포는 다음과 같다.

$$\textbf{Joint Sampling Density} \quad p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{x_n}^T\mathbf{w},\beta^{-1})$$

  pdf를 안다는 것의 의미는 데이터가 가지고 있는 모든 불확실성에 대한 정보를 모조리 다 꿰고 있다는 것. (사실 이것만으로도 굉장히 무지막지하게 큰 가정이다. CLT가 감동인 것은 이런 말도 안 되는 가정이 "사실 우주의 법칙이었다"는 것을 보여주었기 때문.)

- $p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)$에서 $\mathbf{X}$는 주어진 상수로 가정했다. $\mathbf{w}$, β를 알고 있으면 $\mathbf{t}$에 대한 pdf이지만, 거꾸로 생각해서 $\mathbf{t}$를 알고 있다면 이는 모수의 특정한 값 ($\mathbf{w}$, β)이 참일 때 주어진 데이터가 얼마나 "말이 되는지"를 알려주는 **Likelihood** 함수로 볼 수 있다.

- **Maximum Likelihood Principle:** 때문에 **Likelihood는 데이터마다 함수 형태가 다르다!** 그러나 데이터가 무수히 많아지면 결국 Likelihood는 참 모수의 값에서 극대화된다. 때문에 주어진 데이터로 그린 Likelihood를 최대화하는 지점 ($\mathbf{w}$, β)을 모수의 추정치로 삼을 수 있다.

CURVE FITTING IN PROBABILISTIC PERSPECTIVE

- 최적화 문제의 장점은 목적함수에 단조변화함수를 맘껏 취해줄 수 있다는 것이다. 때문에 로그를 취해 $\prod$를 $\sum$으로 바꿔주면

$$\begin{aligned}\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) &= -\frac{\beta}{2}\sum_{n=1}^{N}\{\mathbf{x_n^T w} - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi \\ &= -\frac{\beta}{2}\|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi\end{aligned}$$

- β의 값은 $\mathbf{w}$가 최소화되는 지점에는 영향을 미치지 않는다. 때문에 아까 본 error function을 최소화하는 문제와 똑같다. $\mathbf{w}_{ML} = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{t}$, β에 대해 미분하면 $\beta_{ML}^{-1} = \frac{1}{N}\|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2$
- 이렇게 구한 추정치를 원래의 Likelihood에 넣으면, 우리는 새로운 데이터 t에 대한 predictive distribution, 일종의 확률모델을 얻는다. 이거 돌려서 예측하는 것.

Predictive Distribution $\quad p(t|x, \mathbf{w}_{ML}, \beta_{ML}) = \mathcal{N}(t|\mathbf{x}^T\mathbf{w}_{ML}, \beta_{ML}^{-1})$

Table of Contents

BAYESIAN PROBABILITIES

- 확률의 빈도론적 정의는 확률시행을 무수히 반복할 때의 빈도, "long-run frequency"이다. 근데 이렇게 정의해버리면 "내일 비가 올 확률", "올해 여자친구가 생길 확률" 이런거는 확률이란 말을 쓰기가 애매해진다. 일평생 내일은 단 한번만 오고, 올해 연애 시도를 해봐야 뭐 한 번은 하겠나. 확률보다는 '믿음'이 더 맞겠다.
- 그러나 믿음, belief도 공교롭게도 확률의 세 가지 공리로 표현할 수 있다. 즉 사건의 발생할 확률을 그 사건이 발생할 것이라는 나의 믿음의 강도로 볼 수 있으며, 이런 식으로 불확실성을 직접적으로 정량화할 수 있다. 여기에다가 **Bayes Theorem**을 활용하면 이 믿음을 업데이트할 수 있다!
- 이는 우리의 직관과 상당히 일치한다. 올해 연애나 할 수 있겠어 하고 있는데, 어쩌다가 호감이 있는 상대와 데이트를 한 번 했다고 하자. 그렇다면 연애에 대한 희망이 생기지 않는가? 심지어 대여섯 번 더 만났다고 하자. 희망이 확신이 된다! 베이지언은 이 직관을 고스란히 반영한다.
  - 빈도론은 직관에 호소하기 위해 큰 고생을 해야한다. 통입에서 신뢰구간을 제대로 이해하는 수강생이 있기나 한가? 어려워서가 아니라 애초에 말이 안 돼서 그렇다. 현실에서는 표본평균이 오직 하나만 있기 때문이다. "수 만개의 평행우주에서 구한 표본평균들"과 같은 SF적 개념을 가져와야 이해를 할 수 있다.

BAYESIAN PROBABILITY

Bayes Theorem

- 베이즈 정리 자체는 product rule과 조건부 분포의 정의를 알면 바로 나온다.

$$p(C|E) = \frac{p(C,E)}{p(E)} = \frac{p(C)p(E|C)}{p(E)} = \frac{p(C)p(E|C)}{\sum_{C'} p(C')p(E|C')}$$

- 여기서 분모의 $p(E)$ 개별 C에 의존하지 않는 상수이다. 때문에 다음과 같이 쓰기도 한다.

$$p(C|E) \propto p(C)p(E|C)$$

  어떤 사건 E가 발생했을 때 그 원인이 C일 확률은, 애초에 C가 발생활 확률과, 그 C가 발생했을 때 E의 확률의 곱에 비례한다는 것. 만일 $p(C)$만 알고 있으면 **"사건 발생 후 원인의 확률을 묻는 문제"가 "원인이 주어졌을 때 사건의 확률를 묻는 문제"로 바뀐다는 것.**

- 인과관계가 역전된 것이 보이나? 이 때문에 이를 Inverse Probability라고도 한다. 20세기까지만 해도 대부분의 통계학자는 베이지언을 "불경한 것"으로 혐오했다. 왜 그들은 베이지안을 싫어했을까?

BAYESIAN PROBABILITY

Bayes Theorem: Example

- 올해 연애를 할 확률(믿음)은 $p(Luv) = 0.1$, 못 할 확률은 $p(Sad) = 0.9$라고 하자. 올해 연애를 한다면 그 전에 데이트를 좀 할 것이다. 때문에 $p(Date|Luv) = 0.8$, $p(Date|Sad) = 0.3$이라고 하자(둘이 합쳐서 1이 안된다? 조건이 다르면 다른 분포니까!). 어쩌다보니 눈이 마주친 그대와 데이트를 했다. 그렇다면 내가 올해 연애할 확률은 어떻게 됐을까?

$$p(Luv|Date) = \frac{p(Date|Luv)p(Luv)}{p(Date|Luv)p(Luv) + p(Date|Sad)p(Sad)} = \frac{0.8 \times 0.1}{0.8 \times 0.1 + 0.3 \times 0.9} \approx 0.2$$

- 데이트를 한 번 더 했다면 어떻게 될까? 이 경우 $p(Luv) = 0.2$이다. 똑같은 방식으로

$$p(Luv|Date) = \frac{p(Date|Luv)p(Luv)}{p(Date|Luv)p(Luv) + p(Date|Sad)p(Sad)} = \frac{0.8 \times 0.2}{0.8 \times 0.2 + 0.3 \times 0.8} = 0.4$$

공고해지는 믿음을 보아라. 따스한 한 해가 만들어지고 있다.

BAYESIAN PROBABILITY

Bayesian Inference

- 모수 추정의 문제에서 빈도론적 추론과 베이지언 추론을 비교해보자. $X \sim p(x|\theta)$에서 θ를 추정하는 문제다. 표본 x가 주어졌을 때 $p(x|\theta)$는 Likelihood로 볼 수 있음을 배웠다.
- **Frequentist MLE:**
  주어진 Likelihood를 최대화하는 θ를 추정량으로 삼는다.

$$\theta_{ML} = \arg\max_{\theta} p(x|\theta)$$

  이는 근본적으로 **데이터가 주어졌을 때 데이터가 나올 확률** $p(x|\theta)$**을 극대화**하는 방법이다. 뭐 극한에서는 말이 되긴 한다. 그러나 제한된 표본에서는? 동전 3개 던져서 다 앞면 나오면 뒷면이 나올 확률은 0인가? **MLE는 태생적으로 Overfitting을 하게 된다.**
  - 빈도통계학에서 MLE가 정말 중요한데, 왜냐면 MLE는 CLT처럼 그 극한 분포가 알려져 있기 때문이다. 때문에 추정량의 극한 분포로 신뢰구간, 가설검정 등의 빈도론적 추론을 할 수가 있다. 그러나 머신러닝에서 다루는 대부분의 문제는 $n >> p$가 아니다. 때문에 MLE를 썼다가는 과적화가 되기 마련. 때문에 이를 보완하는 다양한 방법이 있다.

BAYESIAN PROBABILITY

Bayesian Inference

- **Bayesian MAP**:
  좀 더 자연스러운 생각은 **데이터가 주어졌을 때 모수의 확률** $p(\theta|x)$**을 극대화**하는 것이다.

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|x)$$

  모수의 확률? 18~20세기 주류 통계학계가 거품을 문 포인트다. 아니 감히 전지전능한 하나님만이 아는 자연의 섭리를, 미천한 인간은 모르지만 엄연히 존재하는, 고정된 모수를 감히 "확률변수"로 취급하다니? 여기서 베이지안과 빈도론자의 철학에 큰 차이가 나타난다.
- 데이터의 생성에 대한 Sampling Density, Likelihood를 $p(x|\theta)$로 본다고 하자.
- **빈도론자:** 모수 θ는 "unknown but certain". 모르지만 고정된 상수이다. 데이터는 하나밖에 없지만 수만 개의 평행우주에는 똑같은 분포를 따르는 수만 개의 다른 데이터가 있을 것이다. 그 데이터들 모두를 가장 잘 설명하는 하나의 최대점이 참 모수값이다. 그러니 지금 가지고 있는 하나의 데이터만을 가장 잘 설명하는 추정치 θ_{ML}를 써도 되지 않겠냐!

BAYESIAN PROBABILITY

Bayesian Inference

- **베이지언:** 모수 θ는 "unknown thus uncertain". **모르니까 확률변수야!** 하나님만이 아는 정답같은 건 잘 모르겠고 내가 그 정답에 대해 얼마나 잘 모르냐는 알겠다. 그러니 나는 θ에 대한 나의 믿음을 확률로 표현할거다. 내 맘이다.
  $p(\theta)$는 데이터를 보기 전(prior) 나의 믿음이고, $p(\theta|x)$는 데이터를 보고 난 후(posterior)의 나의 믿음이다. 그렇다면 $p(\theta|x)$ 를 어떻게 얻는가? 베이즈 정리를 사용한다!

$$
\begin{aligned}
p(\theta) &\quad \text{Belief in each value of } \theta \text{ prior to data} \\
p(x|\theta) &\quad \text{Likelihood of the data per each value of } \theta \\
p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int_\theta p(x|\theta)p(\theta)d\theta} &\quad \text{Belief in each value of } \theta \text{ posterior to data}
\end{aligned}
$$

베이지안은 모든 통계분석을 베이즈정리로 한다. 베이즈정리가 알파이자 오메가이다!

BAYESIAN PROBABILITY

Bayesian Inference

이전에 든 올해 연애 예시를 들어보자. 이 경우 표본은 데이트 여부이며, 모수는 올해 연애를 할 여부이다. **연애를 하는지 안 하는지에 따라 데이트의 분포가 달라진다.** 즉

$$\textbf{Sampling Density: } \begin{cases} p(Date|\theta = 1) & = 0.8 \\ p(Date|\theta = 0) & = 0.3 \end{cases}$$

- 먼저 빈도론자처럼 생각해보자. θ는 0 혹은 1 둘 중 하나이며, 그 값은 운명의 세 여신 모리아이 자매만 알고 있다. 미천한 인간은 데이트 한 번 하고 나서 이 θ가 0인지 1인지를 결정해야 한다. MLE 원칙에 충실한 빈도론자는 "0.8이 0.3보다 크네"하고 올해 연애를 한다고 결론을 내린다.

$$\therefore \ \theta_{ML} = 1$$

BAYESIAN PROBABILITY

Bayesian Inference

$$\textbf{Sampling Density: } \begin{cases} p(Date|\theta = 1) & = 0.8 \\ p(Date|\theta = 0) & = 0.3 \end{cases}$$

- 베이지안은 이렇게 말한다. **애초에 너가 연애를 할 확률이 굉장히 낮지 않을까?** 아니 뭐 물론 올해 연애한다면 데이트는 당연히 하겠지. 그렇지만 어쩌다 한번 데이트한거 가지고 설레발치는게 아닐까?
- 즉 만일 θ의 값을 하나 골라야한다면, Likelihood 뿐만이 아니라 Prior도 고려해야 한다는 것이다. 이것이 MAP이다. 베이즈 정리에서 분모는 θ의 값에 상관없이 똑같다. 때문에 분자만 고려해보면,

$$\begin{cases} p(Date|\theta = 1)p(\theta = 1) & = 0.8 \times 0.1 = 0.08 \\ p(Date|\theta = 0)p(\theta = 0) & = 0.3 \times 0.9 = 0.27 \end{cases}$$

$$\therefore \ \theta_{MAP} = 1$$

BAYESIAN PROBABILITY

Bayesian Inference

- 하지만 찐 베이지언은 MAP를 하지 않는다. 사실 MLE나 MAP나 똑같다. 전자는 Likelihood라는 함수의 최댓값을 뽑는 거고, 후자는 Likelihood와 Prior까지 같이 고려해 최댓값을 뽑는거니, 결국은 하나의 θ 추정치를 쓴다는 것에서는 똑같다.
- 그러나 베이지안의 철학은 모수도 확률변수라는 게 아닌가! 확률변수를 감히 하나의 값으로 표현할 수 있는가? 확률변수를 표현하는 가장 완전한 방법은 그 분포를 온전히 그려내는 것이다! 즉

$$\textbf{Posterior Belief in } \theta \begin{cases} p(Date|\theta=1)p(\theta=1)/p(Date) & = 0.8 \times 0.1 \approx 0.2286 \\ p(Date|\theta=0)p(\theta=0)/p(Date) & = 0.3 \times 0.9 \approx 0.7714 \end{cases}$$

이 분포를 드러내기 위해 평균 0.22을 쓸 수도 있고, 극빈값 0을 쓸 수도 있다. θ가 연속일 경우 95% **확률**구간을 쓸 수도 있다. 중요한 것은 θ에 내재한 불확실성 구조를 그대로 가져간다는 것!

BAYESIAN PROBABILITY

Bayesian Inference for the Binomial(LAB)

- Beta-Binomial case

PRIOR $\theta \sim \text{beta}(a, b)$

LIKELIHOOD $Y|\theta \sim \text{binomial}(n, \theta)$

POSTERIOR

$$\begin{aligned} p(\theta|y) &= \frac{p(\theta)p(y|\theta)}{p(y)} \\ &= \frac{1}{p(y)} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \times \binom{n}{y}\theta^{y}(1-\theta)^{n-y} \\ &= c(n, y, a, b) \times \theta^{a+y-1}(1-\theta)^{b+n-y-1} \\ &= \text{dbeta}(\theta, a+y, b+n-y)\,. \end{aligned}$$

BAYESIAN PROBABILITY

Bayesian Inference for the Gaussian

- 평균 μ, 분산 λ^{-1}인 정규분포를 따르는 확률변수 $x_1, x_2, ...$를 생각해보자.

$$\textbf{Joint Sampling Density} \quad p(\mathbf{x}|\mu, \lambda) = \prod_{i=1}^{N} (\frac{\lambda}{2\pi})^{\frac{1}{2}} \exp(\frac{\lambda}{2}(x_i - \mu)^2)$$

- 이 함수의 로그를 취해 각각 모수에 대해 편미분하면 MLE 추정량을 얻게된다;

$$\mu_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\lambda_{ML}^{-1} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

BAYESIAN PROBABILITY

Bayesian Inference for the Gaussian unknown μ

$$\textbf{Joint Sampling Density} \quad p(\mathbf{x}|\mu,\lambda) = \prod_{i=1}^{N} (\frac{\lambda}{2\pi})^{\frac{1}{2}} \exp(\frac{\lambda}{2}(x_i-\mu)^2)$$

- 베이지안의 핵심은 $p(\theta|\mathbf{x})$를 구할 때 하는 어마무시한 적분이다. 적분을 손으로 할 수 있으려면 Likelihood와 Prior의 함수 형태가 같아야하는데, 이를 Conjugacy라고 한다.
- 먼저 분산을 이미 알고 평균에 대해 추정한다고 하자. 그렇다면 평균의 분포함수도 지수함수 안에 2차식이 있어야한다. 때문에 μ에 대한 나의 사전 믿음을 정규분포로 표현한다.

$$\textbf{Prior} \quad p(\mu|\mu_0,\lambda_0) = (\frac{\lambda_0}{2\pi})^{\frac{1}{2}} \exp(-\frac{\lambda_0}{2}(\mu-\mu_0)^2)$$

BAYESIAN PROBABILITY

Bayesian Inference for the Gaussian unknown μ

- Posterior 분포도 결국 정규분포를 따를 것을 알면, exp 안의 이차항에서 μ에 대한 계수를 비교함으로써 그 모수를 쉽게 알 수 있다. 즉

$$p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_N, \lambda_N) \propto \exp(-\frac{\lambda_N}{2}\sum(\mu - \mu_N)^2)$$
$$= \exp(-\frac{\lambda_N}{2}\mu^2 + \lambda_n\mu_n\mu + constant)$$

$$p(\mu|\mathbf{x}) \propto p(\mu)p(\mathbf{x}|\mu)$$
$$\propto \exp(-\frac{\lambda}{2}\sum(x_i - \mu)^2 - \frac{\lambda}{2}(\mu - \mu_0)^2)$$

- 따라서 μ의 사후분포의 평균과 분산은 $\lambda_N = \lambda_0 + \lambda$, $\mu_N = \frac{\lambda_0}{\lambda_0 + \lambda}\mu_0 + \frac{\lambda}{\lambda_0 + \lambda}\mu_{ML}$

BAYESIAN PROBABILITY

Bayesian Inference for the Gaussian unknown λ

- 평균을 알고 분산을 모를때, λ에 대해 추론해보자.

$$\textbf{Likelihood: } p(\mathbf{x}|\lambda) = \prod_{i=1}^{N} (\frac{\lambda}{2\pi})^{\frac{1}{2}} \exp(-\frac{\lambda}{2}(x-\mu)^2)$$

$$\propto \lambda^{N/2} \exp(-\frac{\lambda}{2} \sum_{i=1}^{N} (x_i - \mu)^2)$$

이 형태가 익숙하지 않은가? 감마분포다! 사전분포가 감마분포이면 사후분포도 감마분포가 될 것!

- $shape = a, rate = b$인 감마분포의 평균은 a/b, 분산은 a/b^2, mode는 $(a-1)/b$ 이다.

$$\Gamma(\lambda|a,b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda)$$

BAYESIAN PROBABILITY

Bayesian Inference for the Gaussian unknown λ

- 감마분포에 대한 짧은 단상.
  $X_{(0,t)}$는 t 기까지 어떤 사건의 발생 횟수이며, 이는 포아송 프로세스를 따른다고 하자. 또한 T_a을 그 사건이 a번 발생하기까자의 소요시간으로 하자. 그러면 두 확률변수 간에는 다음의 관계가 성립한다. (김해경, 박경옥, 2009)

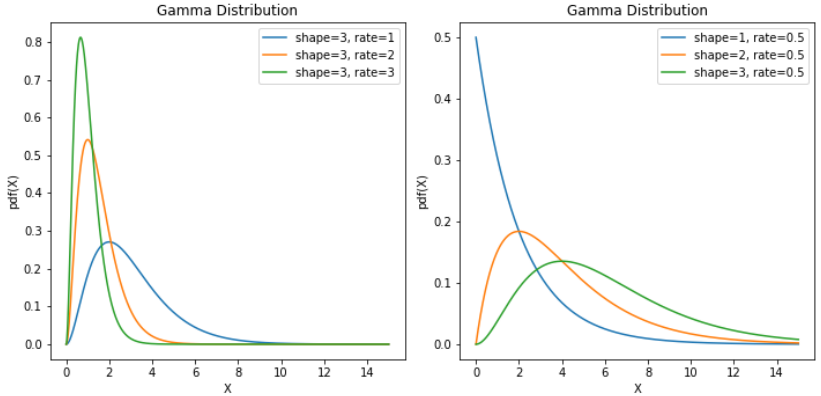
$$
\begin{aligned}
X_{(0,t)} &\sim poi(m = bt) \\
T_1 &\sim \exp(b) = \Gamma(1, b) \\
T_a &\sim \Gamma(a, b)
\end{aligned}
$$

1. 감마 확률변수는 소요시간이므로 0보다 크다.
2. a가 같을 때 b가 클수록 사건이 활발히 일어나는 거니 소요 시간이 짧을 것이다.
3. b가 같을 때 a가 클수록 더 많은 사건이 일어나기까지의 시간이니 소요 시간이 길 것이다.

BAYESIAN PROBABILITY

Bayesian Inference for the Gaussian unknown λ

- 감마분포에 대한 짧은 단상. 모수가 아예 평균과 분산으로 주어지는 정규분포보다 직관적으로 개형을 떠올리기 힘들지만 이렇게 생각해보면 도움이 된다.



BAYESIAN PROBABILITY

Bayesian Inference for the Gaussian unknown λ

$$\textbf{Prior} \quad p(\lambda) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda)$$

$$\textbf{Likelihood} \quad p(\mathbf{x}|\lambda) \propto \lambda^{N/2}\exp(-\frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2)$$

$$\textbf{Posterior} \quad p(\lambda|\mathbf{x}) = \frac{b_N^{a_N}}{\Gamma(a_N)}\lambda^{a_N-1}\exp(-b_N\lambda) \propto \lambda^{a_N-1}\exp(-b_N\lambda)$$

- 따라서 얘도 계수를 비교해보면

$$a_N = a_0 + \frac{N}{2}$$

$$b_N = b_0 + \frac{N}{2}\sigma_{ML}^2 = a_0\frac{b_0}{a_0} + \frac{N}{2}\sigma_{ML}^2$$

여기서 $2a_0$은 사전분포의 표본개수(믿음의 강도), b_0/a_0은 내 마음 속 x_i의 분산이다.

BAYESIAN PROBABILITY

Bayesian Inference for the Gaussian unknown λ

정규분포를 따르는 데이터 x 하나만 고려해보자. 이 정규분포의 평균 μ는 이미 알고 있다. 관심은 τ인데, 감마분포를 따른다. Prior와 Likelihood을 곱한 식은 μ와 τ의 결합확률분포이겠다.

$$\begin{aligned}
\textbf{Prior} \quad p(\tau|a,b) &= \frac{b^a}{\Gamma(a)}\tau^{a-1}\exp(-b\tau) \\
\textbf{Likelihood} \quad p(x|\tau) &= (\frac{\tau}{2\pi})^{\frac{1}{2}}\exp(-\frac{\tau}{2}(x-\mu)^2) \\
\textbf{Posterior} \quad p(\tau|x) &\propto p(x,\tau) = \Gamma(\tau|a,b)\mathcal{N}(x|\tau^{-1}) \\
&= \frac{b^a}{\Gamma(a)}\tau^{a-1}\exp(-b\tau)(\frac{\tau}{2\pi})^{\frac{1}{2}}\exp(-\frac{\tau}{2}(x-\mu)^2)
\end{aligned}$$

그렇다면 이를 τ에 대해 적분하면 어떻게 될까? **t-분포**가 나온다!

BAYESIAN PROBABILITY

Bayesian Inference for the Gaussian unknown λ

$$
\begin{aligned}
p(x) &= \int p(x,\tau)d\tau = \int \Gamma(\tau|a,b)\mathcal{N}(x|\tau^{-1})d\tau \\
&= \int \frac{b^a}{\Gamma(a)}\tau^{a-1}\exp(-b\tau)(\frac{\tau}{2\pi})^{\frac{1}{2}}\exp(-\frac{\tau}{2}(x-\mu)^2)d\tau \\
&= \frac{b^a}{\Gamma(a)}\frac{1}{2\pi}^{1/2}\int \tau^{a+1/2-1}\exp(-\tau(\frac{(x-\mu)^2}{2}+b))d\tau
\end{aligned}
$$

(적분 기호가 있는 식의 값은 $\Gamma(a+1/2, b+\frac{(x-\mu)^2}{2})$ pdf의 상수부분의 역수)

$$
\begin{aligned}
&= \frac{b^a}{\Gamma(a)}\frac{1}{2\pi}^{1/2}\Gamma(a+1/2)[b+\frac{(x-\mu)^2}{2}]^{-a-1/2} \\
\therefore St(x|\mu,\lambda,\nu) &= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}(\frac{\lambda}{\pi\nu})^{1/2}[1+\frac{\lambda(x-\mu)^2}{\nu}]^{-\nu/2-1/2} \quad (\nu=2a, \lambda=a/b)
\end{aligned}
$$

BAYESIAN PROBABILITY

Bayesian Inference for the Gaussian unknown λ

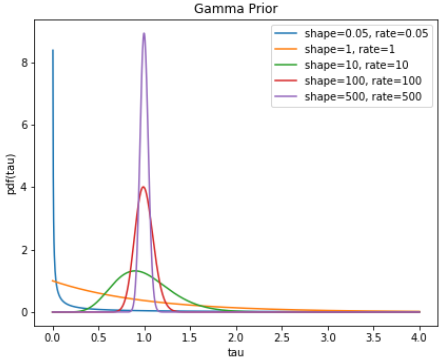
$$St(x|\mu,\lambda,\nu) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}(\frac{\lambda}{\pi\nu})^{1/2}[1+\frac{\lambda(x-\mu)^2}{\nu}]^{-\nu/2-1/2} \quad (\nu = 2a, \lambda = a/b)$$

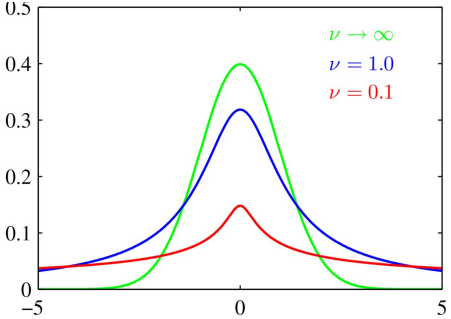
- 베이지안으로 보면 **t 분포는 평균이 μ이지만 분산이 다른 무수히 많은 정규분포의 가중평균**으로 볼 수 있다. 이때 가중치는 분산들, 정확히 말하면 precision τ의 prior 분포 $\Gamma(a,b)$이다. $\nu = 2a$는 사전믿음 형성에서 생각한 표본의 개수이며, $\lambda = a/b$는 그 사전 감마분포의 평균인 셈이다.
- 이때 λ는 그대로인데 ν만 높아지면? 사전 감마분포에서 a와 b가 비율을 유지하면서 같이 극한으로 치닫는 경우다. 이러면 평균은 그대로인채 분산이 0으로 가 결국 뾰족한 degenerate 분포가 된다.
- 그러면 사실상 precision이 λ라는 것에 모두 믿음을 몰빵하는거니, t분포도 결국 $\mathcal{N}(x|\mu,\lambda^{-1})$ 정규분포로 근사하게 된다. 쓸데없이 복잡해보인 t 분포, 그 어떤 다른 설명보다도 더 직관적이지 않은가!

BAYESIAN PROBABILITY

Bayesian Inference for the Gaussian unknown λ

$\nu = 0.1$: 정규분포들인데 대부분 분산이 0인 애들. 그러면 t분포는 0 위주로 뾰족함. (Cauchy distribution). $\nu = 500$: 정규분포들인데 애들 분산이 대부분 0.999아니면 1.001 이런 식. 그거 다 평균하면 그냥 표준정규분포다.





BAYESIAN PROBABILITY

Bayesian Inference for the Gaussian unknown λ, μ

- Prior to Posterior: 4-step ladder process

PRIOR – σ

$$1/\sigma^2 \sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2) \qquad \text{E}[\sigma^2] = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2-1}$$

– σ_0는 "내가 생각하는 모집단의 분산", ν_0는 "그 분산을 계산한 내 마음 속 표본 크기"

PRIOR – μ

$$\theta|\sigma^2 \sim \text{normal}(\mu_0, \sigma^2/\kappa_0)$$

– μ_0는 "내가 생각하는 모집단의 평균", k_0는 "그 평균을 계산한 내 마음 속 표본 크기"
(분산 σ_0인 k_0개의 독립인 확률변수의 합의 평균으로 생각하자)

LIKELIHOOD

$$Y_1, \ldots, Y_n|\theta, \sigma^2 \sim \text{ i.i.d. normal } (\theta, \sigma^2)$$

POSTERIOR

$$p(\theta, \sigma^2|y_1, \ldots, y_n) = p(\theta|\sigma^2, y_1, \ldots, y_n)p(\sigma^2|y_1, \ldots, y_n)$$

MEAN

$$p(\theta|y_1, \ldots, y_n, \sigma^2) \propto p(\theta|\sigma^2)p(y_1, \ldots, y_n|\theta, \sigma^2)$$

VAR

$$p(\sigma^2|y_1, \ldots, y_n) \propto p(\sigma^2)p(y_1, \ldots, y_n|\sigma^2)$$
$$= p(\sigma^2) \int p(y_1, \ldots, y_n|\theta, \sigma^2)p(\theta|\sigma^2)\, d\theta$$

– Prior와 마찬가지로 Joint posterior가 얻어지므로, 평균과 분산 각각에 대한 사후 분포를 구해 따로 추론을 해야 한다.

BAYESIAN PROBABILITY

Bayesian Inference for the Gaussian unknown λ, μ

- Joint posterior for mean and variance

$p(\theta,\sigma^2) = p(\theta|\sigma^2)p(\sigma^2)$

POSTERIOR

$$p(\theta, \sigma^2|y_1, \ldots, y_n) = p(\theta|\sigma^2, y_1, \ldots, y_n)p(\sigma^2|y_1, \ldots, y_n)$$

POST – μ

$$\{\theta|y_1, \ldots, y_n, \sigma^2\} \sim \text{normal}(\mu_n, \sigma^2/\kappa_n)$$

$$\mu_n = \frac{(\kappa_0/\sigma^2)\mu_0 + (n/\sigma^2)\bar{y}}{\kappa_0/\sigma^2 + n/\sigma^2} = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n}$$

$$\kappa_n = \kappa_0 + n$$

– N이 관측치 개수, Y가 확률변수의 합임을 생각하면 k_n는 "내 생각과 실제 관측치의 합",
μ_n는 "내 생각과 실제 관측치를 모두 합산했을 때의 평균", σ^2/k_n은 "전체 관측치 합의 분산"으로 볼 수 있다.

POST– σ

$$\{1/\sigma^2|y_1, \ldots, y_n\} \sim \text{gamma}(\nu_n/2, \nu_n\sigma_n^2/2), \text{ where}$$

$$\nu_n = \nu_0 + n$$

$$\sigma_n^2 = \frac{1}{\nu_n}[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2]$$

– σ_n^2은 prior SSE, sample SSE, 그리고 σ^2추정량의 합을 전체 표본 수로 나눈 것으로 볼 수 있다.

BAYESIAN PROBABILITY

Bayesian Inference for the Gaussian unknown λ, μ

- Example: 날파리 날개 길이
  - Likelihood: 날파리 날개 길이의 분포는 정규분포를 따른다고 가정한다.
  - Prior: 다른 연구에 의하면 날파리 날개 길이의 평균과 분산이 각각 1.9미리와 0.1미리에서 크게 벗어나지 않을 것이라고 한다. 때문에 이러한 믿음에 표본 수 1만큼의 믿음을 부여한다.
    ($\mu_0 = 1.9$ and $\sigma_0^2 = 0.01$, $\kappa_0 = \nu_0 = 1$)
  - Data: 날파리 날개 길이의 표본평균은 1.804이며, 표본 표준편차는 0.13이다.
  - Posterior:

$$\mu_n = \frac{\kappa_0 \mu_0 + n\bar{y}}{\kappa_n} = \frac{1.9 + 9 \times 1.804}{1 + 9} = 1.814$$

$$\sigma_n^2 = \frac{1}{\nu_n}[\nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2] \quad = \frac{0.010 + 0.135 + 0.008}{10} = 0.015\,.$$

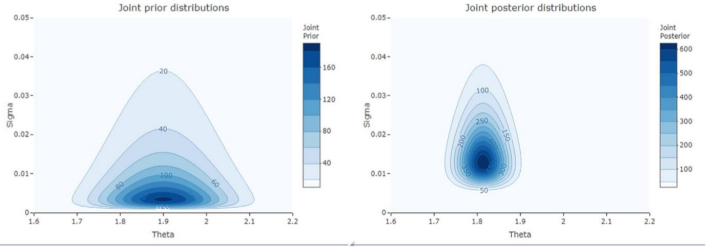


Table of Contents

DECISION THEORY

Classification

- X-ray 사진을 보고 암을 진단하는 경우를 생각해보자. 이는 픽셀 데이터 $\mathbf{x}$를 가지고 $t \in \{C_1, C_2\}$를 결정하는 Binary Classification 문제이다.
- 제한된 데이터를 가지고 전체 불확실성 구조 $p(\mathbf{x}, t)$를 추론하는 것이 Inference인데. 굉장히 어려운 일이다. 그러나 분포를 몰라도 일단 사진을 보고 암인지 아닌지 결정을 하긴 해야하지 않겠나. Decision Theory가 다루는 문제는 이것이다. **확률 분포를 몰라도 어떤 기준에 따라 선택을 하는 것.** 그 기준을 우리는 **loss function**이라고 한다.
- 이 문제의 경우 loss funcion을 잘못될 결정을 내릴 확률로 정의한다.

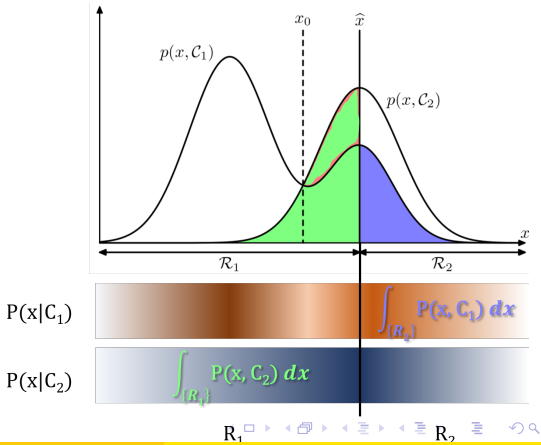
$$\begin{aligned} p(mistake) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} \end{aligned}$$

이걸 최소화하는 $\mathcal{R}_1$, $\mathcal{R}_2$를 "결정"해야 한다.

DECISION THEORY

Classification

- $p(\mathbf{x}|\mathcal{C}_k)$는 $\mathbf{x}$의 범위에서 적분합이 1인 조건부확률분포이다. 위에서 아래로 내려다보면 아래 그라데이션 막대이다.
- 클라스 별 조건부분포에 클라스의 확률(비율)까지 고려하면 결합확률 $p(\mathbf{x}, C_k) = p(C_k)p(\mathbf{x}|C_k)$. 옆의 두 그래프는 조건부가 아닌 결합분포이므로 두 면적을 모두 합하면 1이다.
- 잘못 분류할 확률은 그래프 아래 면적 중에서도 색칠한 부분. 이것을 최소화하려면 **주어진 $\mathbf{x}$ 값에서 결합확률(그래프의 높이)이 가장 높은 C_k로 분류해야 한다.**



DECISION THEORY

Classification

- 맞을 확률을 극대화하는 것으로도 생각할 수 있다. 클라스가 여러 개일 경우

$$p(correct) = \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k) = \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}$$

베이즈 정리를 사용하면

$$= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathcal{C}_k|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

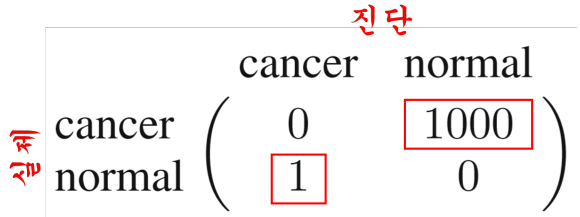
$p(\mathbf{x})$는 클라스와 상관이 없으므로 결국 맞을 확률을 극대화하는 문제는 (loss의 최소화는) 데이터 $\mathbf{x}$를 보고 $p(\mathcal{C}_k|\mathbf{x})$가 가장 높은 클래스에 배정하는 것이다.

- 이 경우 우리는 암묵적으로 암에 걸린 사람의 암을 발견하지 못하는 오류와 멀쩡한 사람을 암 환자로 만드는 오류에 똑같은 가중치를 둔 것.

DECISION THEORY

Classification

- 오류의 가중치가 똑같으면 어떻게 될까? 대부분 정상인보다 암 환자가 훨씬 적다. 오류의 가중치가 같다는 것은 정상인의 오진와 암 환자의 오진을 똑같이 "1건의 오류"로 본다는 것인데, 정상인이 훨씬 더 많으므로 암 판정 기준을 굉장히 엄격하게 잡아, 심하면 모두 정상인으로 판정할 것이다. 예컨대 기준을 한 단계 내리면 암 환자 1명을 잡아도 정상인 10명이 오진이 나니까.
- 때문에 "암 환자의 오진은 정상인의 오진보다 몇 배나 더 심각하다"는 조건을 손실 함수에 반영해야 하는데, 이를 loss matrix라고 한다.



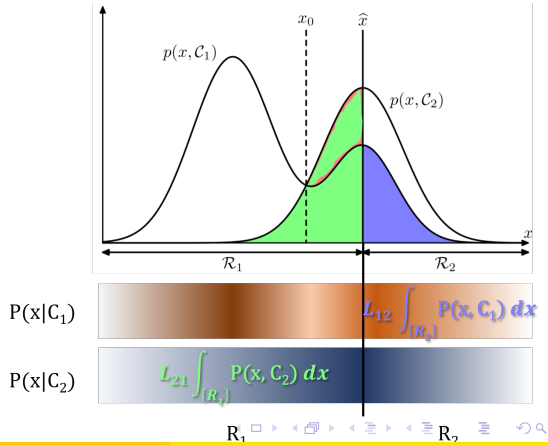
DECISION THEORY

Classification

- L_{kj}는 실제로 k 클라스인 자료를 j로 잘못 분류했을 때의 손실. 이를 고려한 모든 경우에서의 average loss는

$$E[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, C_k) d\mathbf{x}$$

- 베이즈 정리를 이용하면 결국 이를 최소화하는 결정 방법은 데이터 $\mathbf{x}$를 $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$가 최소인 클래스 j에 분류하는 것과 같다. 즉 j **클래스에 분류했을 때 그로 인해 발생하는 오류가 최소이면 그 클래스에 분류하는 것.**



DECISION THEORY

Regression: Squared Loss

- Regression에서 제일 많이 쓰이는 Loss function은 Squared Loss;

$$\textbf{Loss Function:} \quad L(\mathbf{t}, f(\mathbf{x})) = (\mathbf{t} - f(\mathbf{x}))^2$$

$$\textbf{Expected Loss:} \quad E[L] = \int \int (\mathbf{t} - f(\mathbf{x}))^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{t} d\mathbf{x}$$

우리는 많고 많은 $f(\mathbf{x})$ 중에서 이 기대손실을 가장 최소화하는 $f(\mathbf{x})$를 찾고 싶다.

- 이미 수리통계학1 시간에 다음을 알고 있다. 때문에 $f(\mathbf{x}) = E[\mathbf{t}|\mathbf{x}]$로 조건부 기대.

$$E[(y - E[y|x])^2] \leq E[(y - f(x))^2]$$

$$E[(y - f(x))^2] = \underbrace{E[(y - E[y|x])^2]}_{=E[Var(y|x)]} + E[(f(x) - E[y|x])^2]$$

$$= Intrinsic\ variability + Reducible\ Variance$$

DECISION THEORY

Regression: Squared Loss

- 물론 우리는 표본만 보고는 죽었다 깨나도 $E[\mathbf{t}|\mathbf{x}]$를 알지 못한다. 때문에 이를 모분포 $p(\mathbf{x}, \mathbf{t})$를 알때만 구할 수 있다고 해서 Population Minimizer라고 하며, 우리가 데이터로 추정하려고 하는 식이 바로 이거다.
- 변분법을 사용하면 Squared Loss 말고 다양한 Loss function에서 이를 최소화하는 Population Minimizer를 구할 수 있다. (교재 Apx.D에 있긴 한데, 이거만 보고 절대 이해 못하니 **이걸 보자**. 정신건강을 위해선 그냥 skip)

Name	Loss	Derivative	f^*	Algorithm
Squared error	$\frac{1}{2}(y_i - f(\mathbf{x}_i))^2$	$y_i - f(\mathbf{x}_i)$	$\mathbb{E}[y\|\mathbf{x}_i]$	L2Boosting
Absolute error	$\|y_i - f(\mathbf{x}_i)\|$	$\text{sgn}(y_i - f(\mathbf{x}_i))$	$\text{median}(y\|\mathbf{x}_i)$	Gradient boosting
Exponential loss	$\exp(-\tilde{y}_i f(\mathbf{x}_i))$	$-\tilde{y}_i \exp(-\tilde{y}_i f(\mathbf{x}_i))$	$\frac{1}{2}\log\frac{\pi_i}{1-\pi_i}$	AdaBoost
Logloss	$\log(1 + e^{-\tilde{y}_i f_i})$	$y_i - \pi_i$	$\frac{1}{2}\log\frac{\pi_i}{1-\pi_i}$	LogitBoost

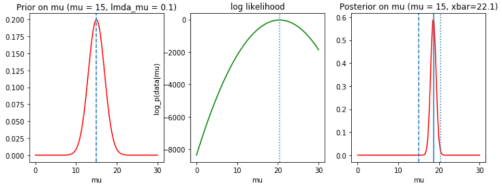
Table 16.1 Some commonly used loss functions, their gradients, their population minimizers f^*, and some algorithms to minimize the loss. For binary classification problems, we assume $\tilde{y}_i \in \{-1, +1\}$, $y_i \in \{0, 1\}$ and $\pi_i = \text{sigm}(2f(\mathbf{x}_i))$. For regression problems, we assume $y_i \in \mathbb{R}$. Adapted from (Hastie et al. 2009, p360) and (Buhlmann and Hothorn 2007, p483).

INFORMATION THEORY

이거는 skip... 나중에 하다가 연관 내용이 나오면 잠깐 소개하는 거로.
MLE가 KL-Divegence를 최소화하는 Esitmator라는 결과가 있는데 범주형자료분석 듣는 사람은 도움될 것.

HOMEWORK

1. Lab1에서 Case 2 해보기. 다음과 같은 그래프를 그려본다.



2. HW: Polynomial Regression에서 다음과 같은 그래프 그려보기 (MSE에 로그 취하세요)