추정 이론

작성자 : 강준하

Main Reference : 머신러닝 이론 입문, 나카이 에츠지

최소제곱법

Training set : $\left\{ \left(x_n,\ t_n\right) \right\}_{n=1}^{10}$

Machine Learning : Training set을 알고리즘에 이용해 x와 t의 함수관계를 추측하는 것

-> 어떤 관측점 x를 새로 관측했을 때 이미 추측한 함수를 사용하여 이 x에 관한 관측값 t를 추정하는 것이 최종 목표이다.

그러면 x와 t 사이에 존재하는 함수관계를 추측해보도록 하자. 다음의 다항식을 보자.

$$f(x) = w_0 + w_1 x + \ldots + w_M x^M = \sum_{m=1}^{M} w_m x^m$$

다항식의 차수인 M의 구체적인 값에 대해서는 나중에 생각하도록 하자. 어쨌든 M이 어떤 값을 가지고 있다면 $M+1$개의 계수 $\left\{w_m\right\}_{m=0}^{M}$가 알 수 없는 parameter로 존재한다. 이들 parameter를 제대로 정해줌으로써 training set을 정확하게 표현하는 다항식을 찾을 수 있다.

어떻게 정확하게 결정할 것인가? 가장 간단한 발상인 최소제곱법을 이용하자. 오차를 다음과 같이 정리하자.

$$\left\{f(x_1) - t_1\right\}^2 + \left\{f(x_2) - t_2\right\}^2 + \ldots + \left\{f(x_{10}) - t_{10}\right\}^2$$

이를 계산의 편의를 위해 반으로 나눈 값을 오차 E_D라고 정의하자. Training set의 개수가 N개라면 오차는 다음과 같다.

$$E_D = \frac{1}{2} \sum_{n=1}^{N} \left\{f(x_n) - t_n\right\}^2$$

이에 $f(x)$를 대입해 정리하면

$$E_D = \frac{1}{2} \sum_{n=1}^{N} \left(\sum_{m=0}^{M} w_m x_n^m - t_n \right)^2$$

와 같고 이를 최소화 시키는 것이 목표이다. 이를 최소화시키기 위해서는 E_D를 $\left\{w_m\right\}_{m=0}^{M}$에 관한 함수로 간주하여 편미분계수가 0이 되도록 하는 조건을 찾아 적용시키면 된다.

$$\frac{\partial E_D}{\partial w_m} = 0 \quad (m = 0,\ \ldots,\ M)$$

(계수를 모두 합쳐서 $\mathbf{w} = (w_0,\ \ldots,\ w_M)^T$ 벡터 형태로 표현하면 기울기 벡터가 0이 된다고 말해도 동치이다. $\nabla E_D(\mathbf{w}) = \mathbf{0}$)

위의 오차 E_D 식의 편미분을 계산하면

$$\sum_{n=1}^{N} \left\{ \sum_{m'=0}^{N} w_{m'} x_n^{m'} - t_n \right\} x_n^m = 0$$ 이고, 이를 정리하여

$$\sum_{m'=0}^{M} w_{m'} \sum_{n=1}^{N} x_n^{m'} x_n^m - \sum_{n=1}^{N} t_n x_n^m = 0$$

를 얻는다.

여기서 x_n^m을 (n, m)원소로 가지는 $N \times (M+1)$행렬 $\mathbf{\Phi}$로 training set을 표현하면 위의 식을 행렬 형식으로 표현 가능하다.

$$\mathbf{w}^T \mathbf{\Phi}^T \mathbf{\Phi} - \mathbf{t}^T \mathbf{\Phi} = \mathbf{0}$$

$\mathbf{w}$는 앞에서 언급한 M차원 벡터 $\mathbf{w} = (w_0, \ldots, w_M)^T$이고, $\mathbf{t}$는 N차원 벡터 $\mathbf{t} = (t_0, \ldots, t_N)^T$이다. 그리고 행렬 $\mathbf{\Phi}$의 원소를 모두 써보면 아래와 같은 형태임을 확인해볼 수 있다.

$$\mathbf{\Phi} = \begin{pmatrix} x_1^0 & x_1^1 & \cdots & x_1^M \\ x_2^0 & x_2^1 & \cdots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ x_N^0 & x_N^1 & \cdots & x_N^M \end{pmatrix}$$

위의 식을 정리하면

$$\mathbf{w} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{t}$$

와 같이 계수 $\mathbf{w}$를 얻을 수 있다. $\mathbf{\Phi}$와 $\mathbf{t}$는 training set에서 정해지는 값이다. 즉 위 식은 training set을 이용해 다항식의 계수 $\mathbf{w}$를 결정할 수 있는 공식이다.

이제 $\mathbf{\Phi}^T \mathbf{\Phi}$의 역행렬의 존재 여부에 대해서 생각해 보자. E_D의 2계 편미분 계수를 나타내는 Hessian을 이용하자. Hessian $\mathbf{H}$는 $(M+1) \times (M+1)$ 인 정방행렬이다. Hessian의 정의에 의해

$$H_{mm'} = \frac{\partial^2 E_D}{\partial w_m \partial w_{m'}} = \sum_{n=1}^{N} x_n^{m'} x_n^m \qquad (m, m' = 0, \ldots, M)$$

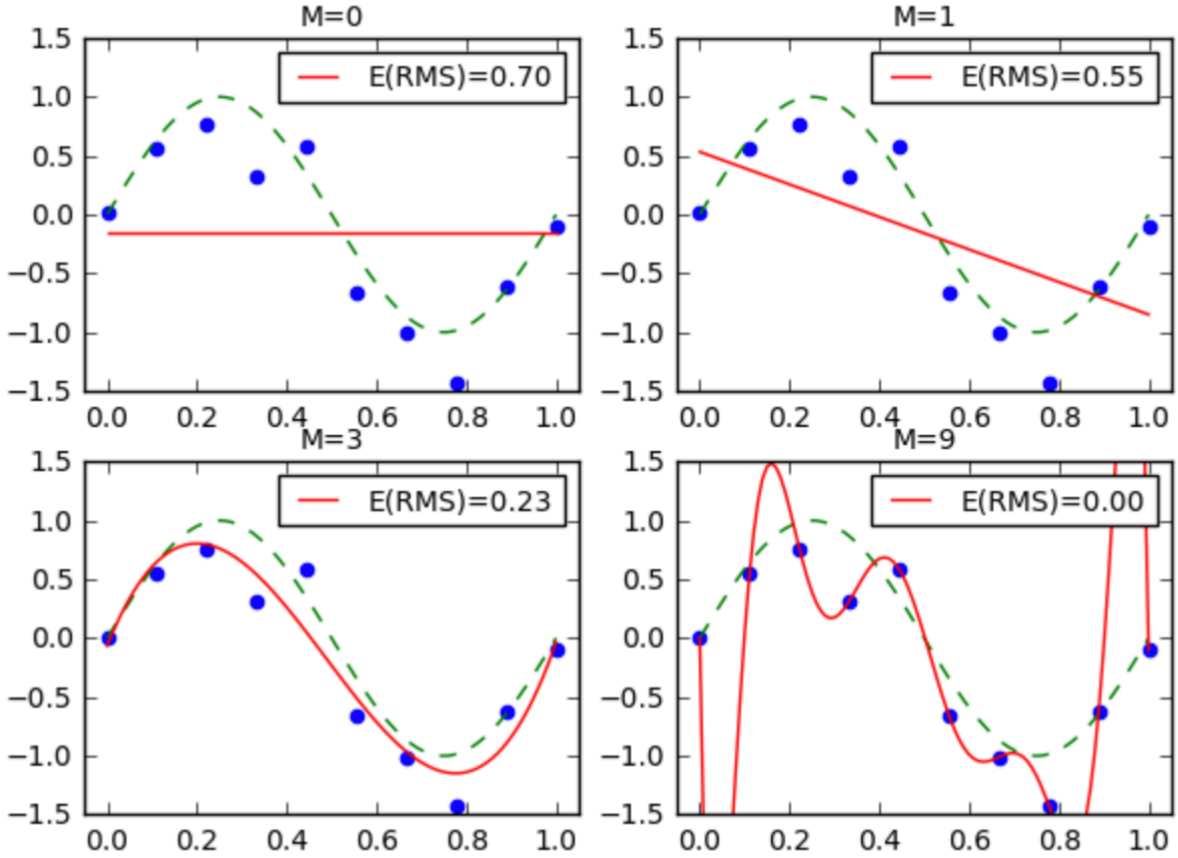
이고, 이를 통해

$$\mathbf{H} = \mathbf{\Phi}^T \mathbf{\Phi}$$

임을 알 수 있다. 그런데, $M + 1 \le N$ 인 경우에는 $\mathbf{\Phi u} \neq 0$ 이므로 항상 다음 부등식

$$\mathbf{u}^T \mathbf{H} \mathbf{u} = \mathbf{u}^T \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{u} = \|\mathbf{\Phi u}\|^2 > 0$$

을 만족한다. 이때 Hessian $\mathbf{H}$를 positive definite matrix(양정치행렬)이라 한다. Positive definite matrix는 역행렬을 가진다는 사실이 이미 증명되어 있으므로(**) $M + 1 \le N$ 인 경우에는 하나의 $\mathbf{w}$를 항상 찾을 수 있다. 그러나 $M + 1 > N$ 인 경우에는 Hessian $\mathbf{H}$가 positive semi-definite matrix(반양정치행렬)이 되어 $(\mathbf{u}^T \mathbf{H} \mathbf{u} \ge 0)$ E_D를 최소로 만드는 $\mathbf{w}$가 여러개 존재하게 된다.



(*) 그래프를 그려서 최종 오차를 확인해 볼 때는 E_D가 아닌 E_{RMS}를 이용한다.

$$E_{RMS} = \sqrt{\frac{2E_D}{N}}$$

이는 평균 제곱근 오차(Root Mean Square Error)라고 부른다. 식의 의미를 해석해본다면 '다항식을 통해 예측할 수 있는 값과 training set 값의 차이를 제곱한 것의 평균값' 이다. 이는 곧 E_{RMS}가 '우리가 다항식을 통해 예상할 수 있는 값과 training set 값들이 평균에서 어느 정도 떨어져 있는지'를 나타낸다는 의미이다.

최대우도법

데이터의 배경에 M차 다항식 관계가 존재하고 표준편차 σ만큼의 오차가 포함되어 있다고 가정해 보자. 표준편차 σ라는 것은 $\pm\sigma$의 범위로 관측 데이터가 변동한다는 의미이다. 최소제곱법에 오차에 관한 가정을 하나 추가한 것이다. 그 다항식은 다음과 같다.

$$f(x) = w_0 + w_1 x + \ldots + w_M x^M = \sum_{m=1}^{M} w_m x^m$$

그리고 관측값 x_n의 관측값 t는 $f(x_n)$을 중심으로 하여 $f(x_n) \pm \sigma$의 범위로 흩어져 있다고 생각하자. μ를 중심으로 하여 $\mu \pm \sigma$의 범위로 흩어지는 난수를 평균이 μ이고 분산이 σ^2인 정규분포로 표현할 수 있다. 이 정규분포는 다음의 그림과 수식으로 표현된다.

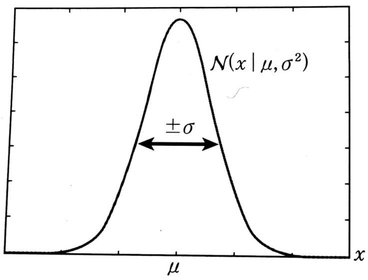


그림 3.2 정규분포의 확률밀도

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

그러나 이 함수는 변수 x값이 난수에 의해 흩어진다는 것이 가정이다. 이번 예제의 경우에는 난수에 의해 흩어지는 것이 관측값 t이므로 그 흩어진 값들의 중심이 $f(x_n)$이다. 위의 함수를 고쳐서 다시 표기하면 다음과 같다.

$$\mathcal{N}(t \mid f(x_n), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\{t - f(x_n)\}^2}$$

이는 다음 그림에서 확인할 수 있듯이 각각의 관측점 x_n에서 관측값 t가 $f(x_n)$을 중심으로 하여 종 모양의 확률로 흩어진다고 생각하면 된다.

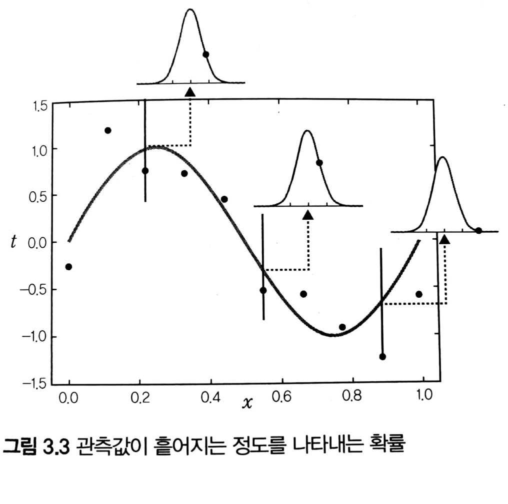


그림 3.3 관측값이 흩어지는 정도를 나타내는 확률

여기서 말하는 관측값 t라는 것은 '이후 새로 관측되는 값'을 의미한다. Training set으로 주어지는 t_n은 이미 관측된 값이고 이후에 새로 관측되는 t는 다른 값이 될 것이다. 이후에 관측되는 값 t의 확률분포가 위의 식으로 계산된다고 생각하면 된다. t_0가 구체적인 값이라고 할 때 $t = t_0$ 값이 얻어질 확률을 알고 싶다면 아래의 식으로 계산하면 된다.

$$\mathcal{N}(t_0 \mid f(x_n), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left\{t_0 - f(x_n)\right\}^2}$$

위 식은 엄밀히 말하면 확률밀도를 나타낸다. Δt를 매우 작은 값이라 생각하고 이후에 얻어질 t값이 $t_0 \sim t_0 + \Delta t$의 범위에 있을 확률밀도가 $\mathcal{N}(t_0 \mid f(x_n), \sigma^2)\Delta t$라는 것이 정확한 표현이다.

이제 위 식을 이용해서 Likelihood function(우도함수)을 만들자. 어떤 관측점 x_n에서 t_n값이 나올 확률은 다음과 같이 나타난다.

$$\mathcal{N}(t_n \mid f(x_n), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left\{t_n - f(x_n)\right\}^2}$$

모든 관측점 $\left\{ \left(x_n, t_n\right) \right\}_{n=1}^{N}$에서 해당 값이 얻어질 확률, 즉 전체적으로 training set $\left\{ \left(x_n, t_n\right) \right\}_{n=1}^{N}$의 데이터가 얻어질 확률 P를 구하려면 각각의 확률을 모두 곱하면 된다.

$$P = \mathcal{N}(t_1 \mid f(x_1), \sigma^2) \times \ldots \times \mathcal{N}(t_N \mid f(x_N), \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid f(x_n), \sigma^2)$$

이 확률은 파라미터 $\left\{w_m\right\}_{m=0}^{M}$와 σ에 의해 값이 변화하므로 이들 파라미터에 관한 함수라고 생각할 수 있다. 이처럼 'training set 데이터가 얻어질 확률'을 파라미터에 관한 함수라고 간주한 것을 likelihood function이라고 부른다.

여기서 "관측된 데이터(training set)는 발생 확률이 가장 높은 데이터임에 틀림없다."라는 가정 하에 확률 P가 최대로 만드는 파라미터를 결정하는 기법을 Maximum Likelihood Estimation(최대우도법)이라고 한다. 이제 P를 최대화시키기 위한 수학적 계산을 시작하자.

P에 대입할 걸 다 대입해서 정리하면 다음과 같은 식을 얻는다.

$$P = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left\{t_n - f(x_n)\right\}^2} = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left\{t_n - f(x_n)\right\}^2\right]$$

마지막 식을 보면 최소제곱법에서 본 E_D를 볼 수 있다.

$$E_D = \frac{1}{2}\sum_{n=1}^{N}\left\{f(x_n) - t_n\right\}^2$$

이를 이용하여 식을 보기좋게 정리하면 다음과 같다.

$$P = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{\sigma^2}E_D}$$

계산의 편의성을 위해 $\beta = \dfrac{1}{\sigma^2}$로 치환하여 정리하면 E_D는 다항식 $\mathbf{w} = (w_0, \ldots, w_M)^T$에 의존하므로 P는 $(\beta, \mathbf{w})$에 의존할 수 있다. 이를 표현하면 다음과 같다.

$$P(\beta, \mathbf{w}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} e^{-\beta E_D(\mathbf{w})}$$

이 식을 최대로 만드는 $(\beta, \mathbf{w})$를 구하면 된다. 편의를 위해 로그를 취하자.

$$\ln P(\beta, \mathbf{w}) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \beta E_D(\mathbf{w})$$

로그함수는 단조증가함수이므로 $\ln P$를 최대로 만드는 것과 P를 최대로 만드는 것은 동일한 의미를 가진다. 이 $\ln P$를 최대화시키는 $(\beta, \mathbf{w})$는 다음의 조건에 의해 결정된다.

$$\frac{\partial(\ln P)}{\partial w_m} = 0 \qquad (m = 0, \ldots, M)$$

$$\frac{\partial(\ln P)}{\partial \beta} = 0$$

위의 식은

$$\frac{\partial E_D}{\partial w_m} = 0 \qquad (m = 0, \ldots, M)$$

과 같이 정리되어 최소제곱법과 동일한 방법을 이용하면 풀 수 있다. 아래의 식은

$$\frac{1}{\beta} = \frac{2E_D}{N}$$

이 되어 표준편차에 관한 식으로 변환하면

$$\sigma = \sqrt{\frac{1}{\beta}} = \sqrt{\frac{2E_D}{N}} = E_{RMS}$$

가 된다. 정리하면 최소제곱법과 동일한 결과를 얻을 수 있다. (굳이 의미부여를 한다면 '다항식을 통해 추정되는 값 $f(x_n)$과 training set 데이터의 평균 오차'를 표준편차 σ의 추정값으로 정한다는 것을 의미한다.)

실제 프로그램을 돌려봐도 유사한 결과를 얻을 수 있다.

베이즈 추정

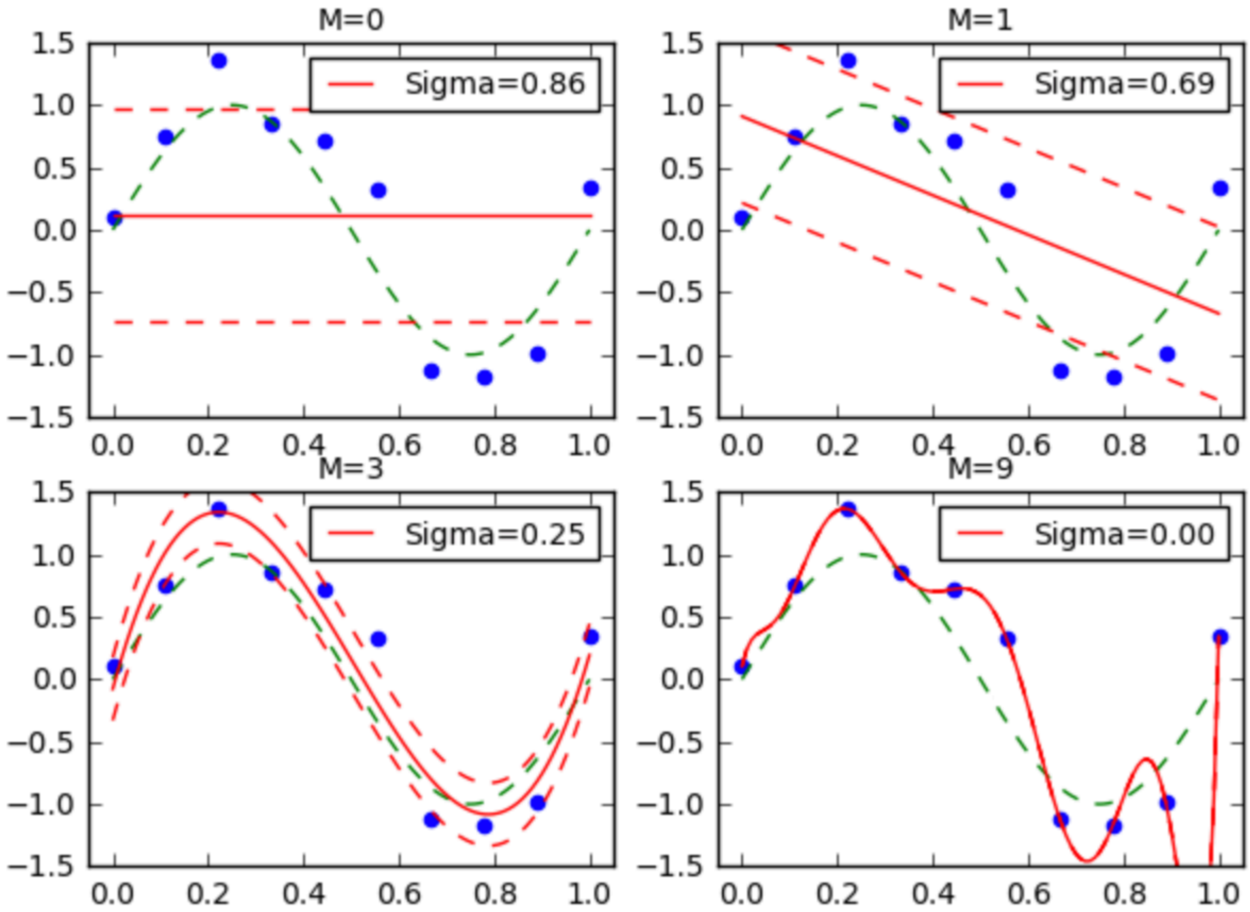
〈Bayes' theorem에 대해서는 알고 있다고 생각하고 설명하지 않겠습니다.〉

위의 maximum likelihood estimation에서 이용한 모델을 이용하여 베이즈 추정을 진행해보자.

$$\mathcal{N}(t \mid f(x_n), \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}\{t - f(x_n)\}^2}$$

$$f(x) = \sum_{m=1}^{M} w_m x^m$$

관측점 x와 관측값 t 사이에는 위 $f(x)$ 형태로 주어진 M차 다항식의 관계가 존재하고 관측값 t는 $f(x)$를 중심으로 하는 분산이 β^{-1}인 정규분포를 따르며 흩어진다고 생각한 것이다. 다항식의 계수 $\{w_m\}_{m=0}^{M}$가 미지의 파라



미터이다. 이들을 $\mathbf{w} = (w_0, \ldots, w_M)^T$로 나타내자. 그리고 계산의 편의를 위해 정규분포의 분산 β^{-1}의 값은 이미 알고 있다고 가정하자.

이제 미지의 파라미터 $\mathbf{w}$에 대한 확률분포를 구성하자. 사전분포 $P(\mathbf{w})$는 어떤 전제조건도 없을 경우의 확률이지만 training set 데이터의 개수 N이 충분히 크다면 임의의 정규분포로 가정해도 된다. 평균이 0이고 분산이 α^{-1}인 정규분포라고 정하자.

$$P(\mathbf{w}) = \mathcal{N}(\mathbf{w}\,|\,\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{\frac{M+1}{2}} \exp\left(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right)$$

이 식이 다변수 정규분포라는데 조심하자. 각 w_m이 평균이 0이고 분산이 α^{-1}인 정규분포를 따른다는 것을 의미한다. ($\mathbf{I}$는 Identity matrix)

그리고 파라미터 $\mathbf{w}$가 정해졌을 경우 training set의 관측값 $\mathbf{t} = (t_0, \ldots, t_N)^T$가 얻어질 확률을 생각해보자. 이는 MLE에서 사용한 식과 동일한 식이다.

$$P(\mathbf{t}\,|\,\mathbf{w}) = \mathcal{N}(t_1\,|\,f(x_1), \beta^{-1}) \times \ldots \times \mathcal{N}(t_N\,|\,f(x_N), \beta^{-1}) = \prod_{n=1}^{N} \mathcal{N}(t_n\,|\,f(x_n), \beta^{-1}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\left[-\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - f(x_n)\}^2\right]$$

이것은 파라미터 $\mathbf{w}$가 정해졌다는 사실을 전제로 한 조건부 확률이다. 이를 토대로 베이즈 정리를 사용하여 사후분포 $P(\mathbf{w}\,|\,\mathbf{t})$를 계산할 수 있다.

$$P(\mathbf{w}\,|\,\mathbf{t}) = \frac{P(\mathbf{t}\,|\,\mathbf{w})}{\int_{-\infty}^{\infty} P(\mathbf{t}\,|\,\mathbf{w}')P(\mathbf{w}')d\mathbf{w}'}P(\mathbf{w})$$

이것은 관측 데이터 $\mathbf{t}$를 기반으로 파라미터 $\mathbf{w}$를 업데이트하는 관계식이다. 분모의 적분식은 다변수에 대한 적분이라 계산이 조금 복잡하지만 $\mathbf{w}$에 종속되지 않으므로 상수 Z로 두고 식을 전개해도 전혀 무리가 없다. 식을 정리해보면 다음과 같다.

$$P(\mathbf{w}\,|\,\mathbf{t}) = Const \times \exp\left[-\frac{\beta}{2}\sum_{n=1}^{N}\left\{f(x_n) - t_n\right\}^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right]$$

이 사후분포 $P(\mathbf{w}\,|\,\mathbf{t})$를 최대로 만드는 $\mathbf{w}$를 결정하면 되므로 지수함수의 안쪽을 최대로 만들면 된다. 이는 곧 아래의 오차함수 E를 최소로 만든다는 조건과 같다.

$$E = \frac{\beta}{2}\sum_{n=1}^{N}\left\{f(x_n) - t_n\right\}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

이 식에서 앞의 항은 최소제곱법에서의 오차함수 E_D와 동일한 형태이다. 따라서 만일 $\alpha = 0$이라면 최소제곱법에서와 동일한 $\mathbf{w}$를 얻을 수 있다. 한편 $\alpha > 0$의 경우 $\mathbf{w}$의 절댓값이 커지면 두 번째 항의 영향으로 오차 E가 커지게 된다. 즉 최소제곱법의 결과보다는 절대값이 작은 $\mathbf{w}$쪽의 확률이 커지게 된다. 이는 사전분포의 영향 때문이다. 평균이 0인 정규분포를 사전분포로 가정했기 때문에 이에 이끌려가는 것으로, 여기서 추정되는 $\mathbf{w}$가 0에 가까워지는 것이다. 이에 α를 줄이면 두 번째 항의 영향을 덜 받게 된다. 이는 사전분포의 분산 α^{-1}이 커짐에 따라 사전분포의 영향이 작아진다는 의미이다.

이 두 번째 항의 의미는 overfitting을 방지하는 데에 있다. 최소제곱법이나 최대우도법에서는 다항식의 계수 M이 커지면 파라미터 $\mathbf{w}$는 다항식 $f(x)$가 모든 training set을 통과하는 overfitting의 늪에 빠지게 된다. 이 overfitting은 파라미터 $\mathbf{w}$가 극단적으로 커져서 발생하는 것이다.

베이즈 추정은 사전분포($P(\mathbf{w}) = \mathcal{N}(\mathbf{w}\,|\,\mathbf{0}, \alpha^{-1}\mathbf{I})$)를 이용하여 $\mathbf{w}$의 절댓값이 그다지 커지지 않도록 억제한다. 이에 다항식 값이 크게 변동하는 것을 막고 overfitting이 발생하지 않도록 조정하는 것이다. 이 억제하는 정도는 α의 크기에 따라 달라지는 것이므로 overfitting을 얼마나 억제하고 싶은 지에 맞춰서 α 값을 조정해야 한다.

이로써 사전분포 $P(\mathbf{w}\,|\,\mathbf{t})$를 최대로 만드는 $\mathbf{w}$가 어떤 값을 갖게 될지 알 수 있게 되었다. 이제 사전분포의 전체적인 형태에 대해서 대략적으로 알아보도록 하자. $P(\mathbf{w}\,|\,\mathbf{t})$에 대한 식을 정리하면 다음과 같은 정규분포가 된다.

$$P(\mathbf{w}\,|\,\mathbf{t}) = \mathcal{N}\left(\mathbf{w}\,|\,\beta\mathbf{S}\sum_{n=1}^{N}t_n\boldsymbol{\phi}(x_n),\ \mathbf{S}\right)$$

다변수 정규분포이므로 분산 $\mathbf{S}$는 행렬의 형태를 갖추고 있다. 분산행렬의 역행렬 $\mathbf{S}^{-1}$이 다음의 식으로 주어진다.

$$\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^T$$

$\phi(x)$는 x를 $0 \sim M$제곱한 값을 나열한 벡터이다.

$$\boldsymbol{\phi}(x) = \begin{bmatrix} x^0 \\ x^1 \\ \vdots \\ x^M \end{bmatrix}$$

파라미터의 사후분포가 정해졌다면 이것을 활용하여 '다음에 관측될 데이터의 확률'을 계산할 수 있다. 파라미터 $\mathbf{w}$가 정해져 있는 상태라면 특정 관측점 x에서 관측값 t가 얻어질 확률은 정규분포 $\mathcal{N}(t \mid f(x), \beta^{-1})$로 주어진다. 이것을 다양한 $\mathbf{w}$에 관해 사후분포 $P(\mathbf{w} \mid \mathbf{t})$라는 가중치를 추가하여 모두 더한다.

$$P(x, t) = \int_{-\infty}^{\infty} P(\mathbf{w} \mid \mathbf{t}) \mathcal{N}(t \mid f(x), \beta^{-1}) d\mathbf{w}$$

이 식에 $P(\mathbf{w} \mid \mathbf{t})$에 대한 식을 대입하면 두 개의 정규분포를 합성하는 적분식이 나온다.

$$P(x, t) = \int_{-\infty}^{\infty} \mathcal{N}\left(\mathbf{w} \mid \beta \mathbf{S} \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n), \mathbf{S}\right) \mathcal{N}(t \mid f(x), \beta^{-1}) d\mathbf{w}$$

일반적으로 이러한 적분식에 적용하는 다음과 같은 공식이 있다.

$$\int_{-\infty}^{\infty} \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}, \mathbf{S}) \mathcal{N}(t \mid \mathbf{a}^T \mathbf{w}, \beta^{-1}) d\mathbf{w} = \mathcal{N}(t \mid \mathbf{a}^T \boldsymbol{\mu}, \beta^{-1} + \mathbf{a}^T \mathbf{S} \mathbf{a})$$

$f(x) = \phi(x)^T \mathbf{w}$에 주의하여 위의 공식을 이용하자.

$$\boldsymbol{\mu} = \beta \mathbf{S} \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n)$$

$$\mathbf{a} = \boldsymbol{\phi}(x)$$

이렇게 대입하면 다음과 같은 정규분포를 얻을 수 있다.

$$P(x, t) = \mathcal{N}(t \mid m(x), s(t))$$

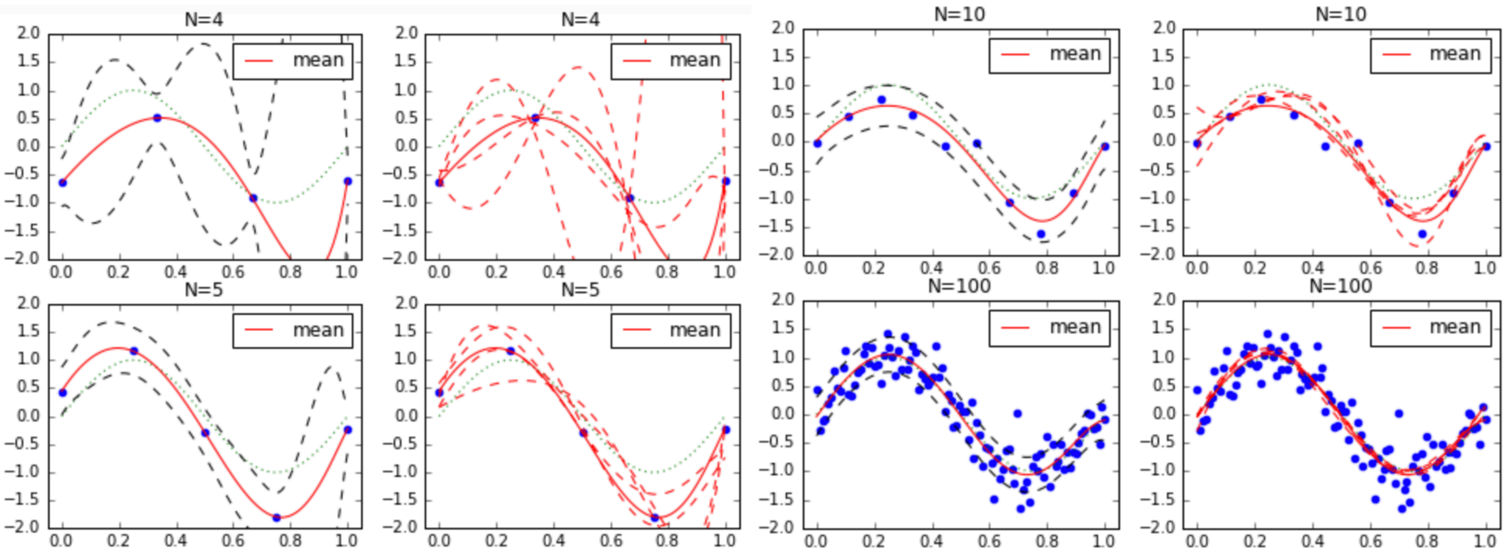
$$m(x) = \beta \boldsymbol{\phi}(x)^T \mathbf{S} \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n)$$

$$s(x) = \beta^{-1} + \boldsymbol{\phi}(x)^T \mathbf{S} \boldsymbol{\phi}(x)$$

이것은 관측점 x를 정하면 그 점에서의 관측 데이터는 평균이 $m(x)$이고 분산이 $s(x)$인 정규분포를 따른다는 간단한 결론을 나타냅니다. $m(x)$과 $s(x)$의 우변에는 training set으로 주어진 데이터 $\left\{ \left(x_n,\ t_n\right) \right\}_{n=1}^{N}$ 이 포함되어 있다는 사실에 주목하자. Training set 데이터를 기반으로 하여 다음에 얻어질 데이터를 추측하는 식이라는 것을 알 수 있다.

추정 코드를 이번에도 돌려서 결과를 확인해보자. 추정에 사용할 다항식의 차수는 $M = 9$이며 사전분포 $P(\mathbf{w})$의 분산은 $\alpha^{-1} = 10000$이라고 정하고 추정을 시작하였다.

그림에 나타난 빨간 실선 그래프는 $y = m(x)$이고, 굵은 점선 그래프는 $y = m(x) \pm s(x)$이다. 얇은 점선 그래프는 본래 데이터의 그래프이다. 이로부터 다음과 같은 사실을 알 수 있다.



- 관측점 개수가 적을 때에는 추정된 평균값이 실제 평균값으로부터 크게 벗어난 부분이 있다. 그러나 그만큼 분산도 크고 실제 평균값은 표준편차 범위 내에 거의 다 들어온다.
- 관측점이 많아지면 표준편차가 작아지고 데이터 개수가 충분히 많다면 난수를 생성하는데 사용한 표준편차 범위 내로 들어온다.
- 사전분포의 영향으로 overfitting이 억제되고 $N = 10$인 경우라도 모든 점을 통과하는 형태가 되지 않는다.