해당 매개 변수의 값. 좀 더 공식적으로 우리는

$$\text{plim}_{\text{미디엄}\to\infty} \hat{\theta}_m = \theta. \tag{5.55}$$

plim 기호는 확률의 수렴을 나타냅니다. $\bullet\bullet 0$,
$P(|\theta_{\text{미디엄}} - \theta| > \bullet\bullet) \to 0$ 같이 $\text{미디엄} \to \infty$. 방정식으로 설명되는 조건 5.55 로 알려져 있습니다 **일관성.** 때로는 약한 일관성이라고도합니다.
강한 일관성은 **거의 확실** 수렴 $\hat{\theta}$ ...에 θ. **거의 확실한 수렴** 일련의 무작위 변수 엑스(1), 엑스(2),. . . 가치에 *엑스*

때 발생 $P($ 임 $_{\text{미디엄}\to\infty}$ 엑스($m) = x) = 1$.

일관성은 데이터 예제 수가 증가함에 따라 추정기에 의해 유도 된 편향이 감소하도록합니다. 그러나 그 반대는 사실이 아닙니다. 점근 적 편향성이 일관성을 의미하지 않습니다. 예를 들어, 평균 모수 추정을 고려하십시오. μ 정규 분포의 *엔(엑스; μ, σ* 2), 다음으로 구성된 데이터 세트 *미디엄* 샘플: { *엑스(* 1), . . . , *엑스(미디엄)}.* 첫 번째 샘플을 사용할 수 있습니다. *엑스(* 1) 데이터 세트의

편향되지 않은 추정기로서 : $\hat{\theta} = x($ 1). 이 경우 이자형($\theta_m) = \theta$ 그래서 견적가
얼마나 많은 데이터 포인트가 표시 되든 편향되지 않습니다. 물론 이것은
추정치가 점근 적으로 편향되지 않습니다. 그러나 이것은 일관된 추정치는 아닙니다. *아니* 그 경우 $\theta_{\text{미디엄}} \to \theta$ 같이 $\text{미디엄} \to \infty$.

5.5 최대 가능성 추정

이전에 우리는 공통 추정기의 몇 가지 정의를보고 그 속성을 분석했습니다. 그러나 이러한 추정치는 어디에서 왔습니까? 어떤 함수가 좋은 추정치를 만들 수 있다고 추측하고 그 편향과 분산을 분석하는 대신, 우리는 다른 모델에 대해 좋은 추정자인 특정 함수를 도출 할 수있는 몇 가지 원리를 원합니다.

가장 일반적인 원칙은 최대 가능성 원칙입니다.

세트를 고려하십시오 *미디엄* 예 X = { *엑스(* 1), . . . , *엑스(미디엄)}* 사실이지만 알려지지 않은 데이터 생성 분포로부터 독립적으로 추출 $P_{\text{데이터}}($ 엑스).

허락하다 $P_{\text{모델}}($ 엑스; $\theta)$ 에 의해 색인 된 동일한 공간에 대한 확률 분포의 매개 변수 군 θ. 다시 말해, $P_{\text{모델}}($ $x; \theta)$ 모든 구성 매핑 *엑스*
실제 확률을 추정하는 실수로 $P_{\text{데이터}}($ *엑스).*

에 대한 최대 가능성 추정량 θ 그런 다음 다음과 같이 정의됩니다.

$$\theta_{\text{ML}} = \underset{\theta}{\text{argmax}}\ P_{\text{미디엄}}\text{odel}(\text{엑스}; \theta) \tag{5.56}$$

$$= \underset{\theta}{\text{argmax}} \prod_{i=1}^{\bullet_{\text{미디엄}}} P_{\text{모델}}(\text{엑스}(\text{나는}); \theta) \tag{5.57}$$

이 제품은 많은 확률에 대해 여러 가지 이유로 불편할 수 있습니다. 예를 들어, 수치 적 흐름이 발생하기 쉽습니다. 더 편리하지만 동등한 최적화 문제를 얻기 위해 우리는

가능성은 변경되지 않습니다 argmax • 그러나 편리하게 제품을 변형합니다
합계로 :

$$\theta_{ML} = \operatorname*{argmax}_{\theta} \sum_{i=1}^{미/디/엄} 로그\ 피/모델(\ 엑스\ (i);\ \theta). \tag{5.58}$$

때문에 argmax 비용 함수를 다시 조정할 때 변경되지 않습니다. *미/디/엄* 기대치로 표현되는 기준의 버전을 얻기 위해
경험적 분포와 관련하여 *피/*데이터 훈련 데이터로 정의 :

$$\theta_{ML} = \operatorname*{argmax}_{\theta} 이자형_{엑스 \sim 피/데이터} 로그\ 피/모델(\ 엑스;\ \theta). \tag{5.59}$$

최대 가능성 추정을 해석하는 한 가지 방법은 최소화하는 것으로 보는 것입니다.
경험적 분포의 비 유사성 *피/*데이터 훈련 세트와 모델 분포에 의해 정의되며, 둘 사이의 비 유사성 정도

KL 발산으로 측정됩니다. KL 발산은 다음과 같이 주어진다.

$$디/KL(\ 피/데이터 \bullet 피/모델) = 이자형_{엑스 \sim 피/데이터} [로그\ 피/데이터(\ 엑스) - 로그\ 피/모델(\ x)]. \tag{5.60}$$

왼쪽의 용어는 모델이 아닌 데이터 생성 프로세스의 기능입니다. 즉, KL 발산을 최소화하기 위해 모델을 훈련 할 때

$$- 이자형_{엑스 \sim 피/데이터} [로그\ 피/모델\ (\ 엑스)] \tag{5.61}$$

물론 방정식의 최대화와 동일합니다. 5.59 .

이 KL 발산을 최소화하는 것은 분포 간의 교차 엔트로피를 최소화하는 것과 정확히 일치합니다. 많은 저자들은"교차 엔트로피"라는 용어를 사용하여 베르누이 또는 소프트 맥스 분포의 음의 로그 우도를 구체적으로 식별하지만 이는 잘못된 이름입니다. 음의 로그 가능 도로 구성된 손실은 훈련 세트로 정의 된 경험적 분포와 모델로 정의 된 확률 분포 사이의 교차 엔트로피입니다. 예를 들어 평균 제곱 오차는 경험적 분포와 가우시안 모델 간의 교차 엔트로피입니다.

따라서 모델 분포를 경험적 분포와 일치시키려는 시도로 최대 가능성을 볼 수 있습니다. *피/*데이터. 이상적으로, 우리는 일치하고 싶습니다
진정한 데이터 생성 분포 *피/*데이터, 하지만 우리는이 배포판에 직접 접근 할 수 없습니다.

최적의 동안 θ 우도를 최대화하든 KL 발산을 최소화하든 상관없이 목적 함수의 값은 동일합니다.

다릅니다. 소프트웨어에서 우리는 종종 두 가지 모두 비용 함수를 최소화한다고 표현합니다. 따라서 최대 가능성은 음의 로그 가능도 (NLL)의 최소화 또는 동등하게 교차 엔트로피의 최소화가됩니다. 최소 KL 발산으로서의 최대 가능성 관점은 KL 발산이 알려진 최소값이 0이기 때문에이 경우에 도움이됩니다. 음의 로그 가능성은 실제로 다음과 같은 경우 음수가 될 수 있습니다. *엑스* 실제 가치입니다.

5.5.1 조건부 로그 가능성과 평균 제곱 오차

최대 우도 추정기는 우리의 목표가 조건부 확률을 추정하는 경우에 쉽게 일반화 될 수 있습니다. *피(* y | 엑스; *θ)* 예측하기 위해 와이
주어진 엑스. 이것은 대부분의지도 학습의 기초를 형성하기 때문에 실제로 가장 일반적인 상황입니다. 만약 *엑스* 우리의 모든 입력을 나타내고 *와이* 관찰 된 모든 목표, 조건부 최대 가능성 추정치는 다음과 같습니다.

$$\theta_{\text{ML}} = \underset{\theta}{\text{argmax}}\ P(Y \mid X; \theta). \tag{5.62}$$

예가 ii로 가정되는 경우 • d. 다음으로 분해 될 수 있습니다.

$$\theta_{\text{ML}} = \underset{\theta}{\text{argmax}} \sum_{i=1}^{m} \log P(y^{(i)} \mid x^{(i)}; \theta). \tag{5.63}$$

예 : 최대 가능성으로서의 선형 회귀 선형 회귀,
섹션 앞부분에서 소개 5.1.4 , 최대 가능성 절차로 정당화 될 수 있습니다. 이전에는 입력을받는 방법을 배우는 알고리즘으로 선형 회귀를 동기 부여했습니다. *엑스* 출력 값을 생성 *와이.* 매핑 *엑스* ...에 *와이* 우리가 어느 정도 임의로 도입 한 기준 인 평균 제곱 오차를 최소화하기 위해 선택됩니다. 이제 최대 가능성 추정의 관점에서 선형 회귀를 다시 살펴 봅니다. 단일 예측을 생성하는 대신 *와이,* 이제 모델을 조건부 분포를 생성하는 것으로 생각합니다. *p (y | 엑스).* 매우 큰 훈련 세트를 사용하면 동일한 입력 값을 가진 여러 훈련 예제를 볼 수 있다고 상상할 수 있습니다. *엑스* 그러나 다른 가치 *와이.* 학습 알고리즘의 목표는 이제 분포를 맞추는 것입니다. *p (y | 엑스)* 그 모든 다른 *와이* 모두 호환되는 값 *엑스.* 이전에 얻은 동일한 선형 회귀 알고리즘을 유도하기 위해 다음을 정의합니다. *p (y | x) = 엔(와이; ŷ (엑스; w), σ*2). 함수 *와이(엑스; w)*

가우스 평균의 예측을 제공합니다. 이 예에서는 분산이 일정한 상수로 고정되어 있다고 가정합니다. *σ*2 사용자가 선택합니다. 우리는 기능적 형태의 선택이 *p (y | 엑스)* 최대 가능성 추정 절차가 이전에 개발 한 것과 동일한 학습 알고리즘을 생성하도록합니다. 이후

예는 조건부 로그 우도 인 iid (방정식 5.63)는 다음과 같이 주어진다.

$$\text{로그 } p\left(y^{(i)} \mid \text{엑스}^{(\text{나는})}; \theta\right) \tag{5.64}$$

$$= -\text{미디엄} \text{로그 } \sigma - \frac{\text{미디엄}}{2} \text{로그 } (2\pi) - \sum_{i=1}^{\text{미디엄}} \frac{\left(\hat{y}^{(\text{나는})} - \text{와이}^{(\text{나는})}\right)^2}{2\sigma^2}, \tag{5.65}$$

어디 $\hat{y}^{(\text{나는})}$에 대한 선형 회귀의 출력입니다. 나는- 일 입력 $\text{엑스}^{(\text{나는})}$과 미디엄 훈련 예제의 수입니다. 로그 우도를 평균 제곱 오차와 비교하면,

$$\text{MSE}_{\text{기차}} = \frac{1}{\text{미디엄}} \sum_{i=1}^{m} \left\| \hat{y}^{(\text{나는})} - \text{와이}^{(i)} \right\|_2, \tag{5.66}$$

우리는 다음에 대해 로그 우도를 최대화하는 것을 즉시 알 수 있습니다. w 모수의 동일한 추정치를 산출합니다. w 평균 제곱 오차를 최소화하는 것과 같습니다. 두 기준은 값이 다르지만 최적의 위치는 동일합니다. 이것은 최대 우도 추정 절차로서 MSE의 사용을 정당화합니다. 보시다시피 최대 우도 추정기는 몇 가지 바람직한 속성을 가지고 있습니다.

5.5.2 최대 가능성의 속성

최대 우도 추정기의 주요 매력은 예의 수와 같이 점근 적으로 최상의 추정기임을 나타낼 수 있다는 것입니다. $\text{미디엄} \to \infty$, 수렴 속도 측면에서 미디엄 증가합니다.

적절한 조건에서 최대 우도 추정기는 일관성 속성을 갖습니다 (섹션 참조). 5.4.5 즉, 훈련 예제의 수가 무한대에 가까워 질수록 매개 변수의 최대 가능성 추정치는 매개 변수의 실제 값으로 수렴됩니다. 이러한 조건은 다음과 같습니다.

- 진정한 분포 $\text{피}_{\text{데이터}}$ 모델 패밀리 내에 있어야합니다. $\text{피}_{\text{모델}}(\cdot; \theta)$.
  그렇지 않으면 추정자가 복구 할 수 없습니다. $\text{피}_{\text{데이터}}$.
- 진정한 분포 $\text{피}_{\text{데이터}}$ 정확히 하나의 값과 일치해야합니다. θ. 다른-
  현명한 최대 가능성은 올바른 $\text{피}_{\text{데이터}}$, 그러나 어떤 값을 결정할 수 없습니다 θ 데이터 생성 처리에 사용되었습니다.

최대 우도 추정기 외에 다른 귀납적 원리가 있으며, 그 중 대부분은 일관된 추정자가되는 속성을 공유합니다. 하나,

일관된 추정자는 자신의 **통계적 효율성,** 하나의 일관된 추정자가 고정 된 수의 샘플에 대해 더 낮은 일반화 오류를 얻을 수 있음을 의미합니다. *미디엄,* 또는 동등하게 고정 된 수준의 일반화 오류를 얻기 위해 더 적은 예제가 필요할 수 있습니다.

통계적 효율성은 일반적으로 **파라 메트릭 케이스 (** 선형 회귀처럼) 우리의 목표는 함수의 값이 아니라 매개 변수의 값을 추정하는 것입니다 (그리고 실제 매개 변수를 식별 할 수 있다고 가정). 실제 모수에 얼마나 가까운지를 측정하는 방법은 예상 평균 제곱 오차에 의해 예상치가 끝났을 때 추정 된 모수 값과 실제 모수 값 사이의 제곱 차이를 계산합니다. *미디엄* 데이터 생성 분포에서 훈련 샘플. 그 모수 평균 제곱 오차는 다음과 같이 감소합니다. *미디엄* 증가하고 *미디엄* large, Cramér-Rao 하한 (라오 , 1945 년 ; 크레이머 , 1946 년)는 일관된 추정기가 최대 가능도 추정기보다 낮은 평균 제곱 오차를 갖지 않음을 보여줍니다.

이러한 이유 (일관성 및 효율성)로 인해 최대 가능성은 종종 기계 학습에 사용하는 선호되는 추정기로 간주됩니다. 과적 합 동작을 생성 할 수있을만큼 예제 수가 적을 때 가중치 감쇄와 같은 정규화 전략을 사용하여 훈련 데이터가 제한 될 때 분산이 적은 편향된 버전의 최대 가능성을 얻을 수 있습니다.

5.6 베이지안 통계

지금까지 우리는 **빈도주의 통계** 단일 값 추정에 기반한 접근 방식 θ, 그런 다음 하나의 추정치를 기반으로 모든 예측을 수행합니다. 또 다른 접근 방식은 가능한 모든 값을 고려하는 것입니다. θ 예측할 때. 후자는 도메인입니다 **베이지안 통계.**

섹션에서 논의한대로 5.4.1 , 빈도주의 관점은 사실입니다
매개 변수 값 θ 고정되었지만 알려지지 않았지만 포인트 추정치는 θ 데이터 세트의 함수이기 때문에 랜덤 변수입니다 (무작위로 표시됨).

통계에 대한 베이지안 관점은 상당히 다릅니다. 베이지안은 확률을 사용하여 지식 상태의 확실성 정도를 반영합니다. 데이터 세트는 직접 관찰되므로 무작위가 아닙니다. 반면에 실제 매개 변수는 θ

알 수 없거나 불확실하므로 무작위 변수로 표시됩니다.

데이터를 관찰하기 전에 우리는 θ 사용 **사전 확률 분포,** $p(\theta)$ (단순히 "이전"이라고도 함). 일반적으로 기계 학습 실무자는 높은 수준의 불확실성을 반영하기 위해 상당히 넓은 (즉, 높은 엔트로피가있는) 사전 분포를 선택합니다.

가치 θ 데이터를 관찰하기 전에. 예를 들어 다음과 같이 가정 할 수 있습니다. *선험적으로* 그 θ 일정한 분포로 일정한 범위 또는 부피에 속합니다. 그 대신 많은 이전의 경우 "단순한"솔루션에 대한 선호도를 반영합니다 (예 : 더 작은 크기 계수 또는 상수에 더 가까운 함수).

이제 데이터 샘플 세트가 있다고 가정합니다. *엑스*(1), . . . , *엑스*(*미디엄*) }. 우리는 다음에 대한 우리의 믿음에 대한 데이터의 효과를 복구 할 수 있습니다. θ 데이터 가능성을 결합하여 *p* (*x* (1), . . . , *엑스*(*m*) | θ) 이전 via Bayes의 규칙 :

$$피(\theta \mid 엑스(1), \ldots, 엑스(m)) = \frac{p(x(1), \ldots, 엑스(m) \mid \theta)\, 피(\theta)}{p(x(1), \ldots, 엑스(미디엄))} \tag{5.67}$$

베이지안 추정이 일반적으로 사용되는 시나리오에서 사전은 상대적으로 균일하거나 높은 엔트로피를 가진 가우스 분포로 시작되며 데이터 관찰은 일반적으로 사후가 엔트로피를 잃고 가능성이 높은 매개 변수 값에 집중하게합니다.

최대 가능성 추정과 비교하여 베이지안 추정은 두 가지 중요한 차이를 제공합니다. 첫째, 포인트 추정치를 사용하여 예측하는 최대 가능성 접근법과는 달리 θ, 베이지안 접근법은 예측을하는 것입니다.

전체 배포를 사용하여 θ. • 예를 들어 관찰 후 *미디엄* 예,
다음 데이터 샘플에 대한 예측 분포, *엑스*(*m* + 1) , ~에 의해 주어진다

$$p(x(m+1) \mid 엑스(1), \ldots, 엑스(m)) = \int p(x(m+1) \mid \theta)\, 피(\theta \mid 엑스(1), \ldots, 엑스(미디엄))\, 디\theta. \tag{5.68}$$

여기에 각 가치 θ 확률 밀도가 양수이면 다음 예의 예측에 기여하며, 기여도는 사후 밀도 자체에 의해 가중됩니다. 관찰 한 후 { *엑스*(1), . . . , *엑스*(*미디엄*)}, 우리가 여전히 가치에 대해 확실하지 않다면 θ, 이 불확실성은 우리가 할 수있는 모든 예측에 직접적으로 통합됩니다.

섹션에서 5.4 , 우리는 빈도 주의적 접근 방식이 불확실성을 어떻게 해결하는지 논의했습니다. 주어진 포인트 추정치의 오염 θ 분산을 평가하여. 추정치의 분산은 관측 된 데이터의 대체 샘플링으로 추정치가 어떻게 변경 될 수 있는지에 대한 평가입니다. 추정기에서 불확실성을 처리하는 방법에 대한 베이지안의 대답은 단순히 그것을 통합하는 것이며, 이는 과적 합으로부터 잘 보호하는 경향이 있습니다. 물론이 적분은 확률 법칙의 적용 일 뿐이며 베이지안 접근 방식을 정당화하기 쉽게 만드는 반면, 추정량을 구성하는 빈도주의 기계는 데이터 세트에 포함 된 모든 지식을 단일 지점으로 요약하는 다소 임시적인 결정에 기반합니다. 견적.

추정에 대한 베이지안 접근법과 최대 가능성 접근법 사이의 두 번째 중요한 차이점은 베이지안의 기여 때문입니다.

사전 배포. 사전은 확률 질량 밀도를 선호하는 매개 변수 공간의 영역으로 이동함으로써 영향을 미칩니다. *선험적으로.* 실제로, 사전은 종종 더 단순하거나 더 부드러운 모델에 대한 선호를 표현합니다. 베이지안 접근 방식의 비평가들은 예측에 영향을 미치는 주관적인 인간 판단의 원천으로 사전을 식별합니다.

베이지안 방법은 일반적으로 제한된 훈련 데이터를 사용할 수있을 때 훨씬 더 잘 일반화되지만 일반적으로 훈련 예제 수가 많을 때 높은 계산 비용으로 인해 더 문제가됩니다.

예 : 베이지안 선형 회귀 여기에서는 선형 회귀 매개 변수를 학습하기위한 베이지안 추정 방법을 고려합니다. 선형 회귀에서는 입력 벡터에서 선형 매핑을 배웁니다. *엑스* ∈ 아르 자형 *엔* 스칼라 값을 예측하려면 *와이* ∈ 아르 자형. 예측은 벡터로 매개 변수화됩니다. *w* ∈ 아르 자형 *엔:*

$$\hat{y} = w \cdot 엑스. \tag{5.69}$$

주어진 세트 *미디엄* 훈련 샘플 (*엑스(* 기차), *와이(* 기차)), 우리는 예측을 표현할 수 있습니다 *와이* 전체 교육 세트에 대해 다음과 같이 설정합니다.

$$와이(\text{기차}) = 엑스(\text{기차})\ w. \tag{5.70}$$

가우스 조건부 분포로 표현됩니다. *와이(* 기차), 우리는

$$피(와이(\text{기차}) \mid 엑스(\text{기차}), w) = 엔(와이 \cdot (\text{기차}); 엑스(\text{기차})\ w, 나) \tag{5.7 • 1}$$

$$\propto 특급 - (와이 \frac{1}{2} (\text{기차}) - 엑스(\text{기차})\ w) \cdot (와이(\text{기차}) - 엑스(\text{기차})\ w), \tag{5.72}$$

가우스 분산이 다음과 같다고 가정 할 때 표준 MSE 공식을 따릅니다. *와이* 하나입니다. 다음에서 표기의 부담을 줄이기 위해
(*엑스(* 기차), *y (* 기차)) 간단히 (*X, y).*

모형 모수 벡터에 대한 사후 분포를 확인하려면 *w,* 먼저 사전 분포를 지정해야합니다. 사전은 이러한 매개 변수의 가치에 대한 우리의 순진한 믿음을 반영해야합니다. 모델의 매개 변수 측면에서 우리의 이전 신념을 표현하는 것이 때때로 어렵거나 부자연 스럽지만, 실제로 우리는 일반적으로 높은 수준의 불확실성을 표현하는 상당히 광범위한 분포를 가정합니다.

약 *θ.* 실수 매개 변수의 경우 • Gau를 사용하는 것이 일반적입니다 • 이전으로 ssian 분포:

$$피(w) = 엔(w; \mu 0, \Lambda 0) \propto 특급 - 1 \frac{1}{2} (w - \mu) \cdot \Lambda 0 (w - \mu 0), \tag{5.73}$$

어디 μ_0 과 Λ_0 각각 사전 분포 평균 벡터와 공분산 행렬입니다. [1]

이렇게 지정된 사전을 사용하여 이제 다음을 결정할 수 있습니다. **후부**
일에 분포 • e 모델 매개 변수.

$$피(w | X, y) \propto 피(y | X, w) 피(w) \tag{5.74}$$

$$\propto 특급 • - (와이 | - Xw) • (와이 | - Xw) 특급 - (w - \mu) • \Lambda - \frac{1}{2} \quad 0 \quad 0 (w •• 0) - \mu \tag{5.75}$$

$$\propto 특급 - 1 - 2 - 와이 | • Xw + w • 엑스 • Xw + w • \Lambda - 1 \quad 0 w - 2 \mu_0 \Lambda_0 w \quad . \tag{5.76}$$

이제 우리는 $\Lambda = 엑스 • X_{미디엄} \Lambda - 1$ $\quad 0$ $\quad$ • $- 1$ 과 $\mu_m = \Lambda 미디엄 엑스$ $\quad$ • $\quad y + \Lambda_0 \quad - 1 \mu_0$. 사용
이 새로운 변수, 우리는 • 사후가 다음과 같이 다시 쓰여질 수 있음을 찾으십시오. • 가우시안
분포:

$$피(w | X, y) \propto 특급 • - (w - \mu_{미디엄}) • \Lambda - 1 \quad 미디엄 (w - \mu 미디엄) • 2 + \frac{1}{2} \mu • \Lambda_{mm} \mu 1 미디엄 \tag{5.77}$$

$$\propto 특급 \quad - 1 - (w - \mu 미디엄) • \Lambda - 1 미디엄 (w - \mu 미디엄) . \tag{5.78}$$

모수 벡터를 포함하지 않는 모든 항 w 생략되었습니다. 이는 분포가 다음에 통합되기 위해
정규화되어야한다는 사실에 의해 암시됩니다. 1.
방정식 3.23 다변량 가우스 분포를 정규화하는 방법을 보여줍니다.

이 사후 분포를 살펴보면
베이지안 추론의 효과. 대부분의 상황에서 우리는 μ ...에 0. 우리가 설정하면 $\Lambda = 1$ $\quad 0$ $\quad$ α 나는,
그때 $\mu_{미디엄}$ 동일한 추정치를 제공합니다 w 가중치 감소 페널티가있는 빈도주의 선형 회귀와 마찬가지로 $\alpha w • w$. 한 가지 차이점은 베이지안 추정치가
정의되지 않은 경우 $\alpha 0$으로 설정됩니다.-우리는 아주 넓은 범위의 사전에 베이지안 학습 프로세스를 시작할 수 없습니다. w. 더 중요한 차이점은 베이지안 추정치가 공분산 행렬을 제공하여

다른 값 w 견적 만 제공하는 것이 아니라 $\mu_{미디엄}$.

5.6.1 최대 *포스터 리 오리* (MAP) 추정

가장 원칙적인 접근 방식은 모수에 대한 전체 베이지안 사후 분포를 사용하여 예측하는 것입니다. θ, 여전히
종종

[1] 특정 공분산 구조를 가정 할 이유가없는 한 일반적으로 대각 공분산 행렬을 가정합니다. $\Lambda_0 = \text{diag} (\lambda_0)$.

단일 포인트 추정. 포인트 추정을 원하는 한 가지 일반적인 이유는 가장 흥미로운 모델에 대해 베이지안 사후를 포함하는 대부분의 연산이 다루기 어렵고 포인트 추정이 다루기 쉬운 근사를 제공하기 때문입니다. 단순히 최대 우도 추정치로 돌아가는 대신, 포인트 추정치의 선택에 영향을 미치기 전에 사전에 허용함으로써 베이지안 접근 방식의 이점을 얻을 수 있습니다. 이를 수행하는 한 가지 합리적인 방법은 **최대 사후 (** MAP) 포인트 추정. MAP 추정치는 최대 사후 확률 (또는 더 일반적인 연속적인 경우의 최대 확률 밀도 지점을 선택합니다. *θ) :*

$$\theta_{\text{지도}} = \operatorname*{argmax}_{\theta} \text{피}(\theta \mid x) = \operatorname*{argmax}_{\theta} \text{로그 피}(x \mid \theta) + \text{로그 피}(\theta). \tag{5.79}$$

위의 오른쪽에서 로그 *피(x | θ),* 즉, 표준 로그 가능도 항 로그 *피(θ),* 사전 배포에 해당합니다.

예를 들어, 가우스가있는 선형 회귀 모델을 고려하십시오.
무게 *w.* 이 사전이 주어진 경우 *엔(w;* 0, 1 $\overline{\lambda}$ *나는* 2), 다음 로그-사전 용어
방정식 5.79 익숙한 것에 비례합니다 *λ w ⋅ w* 체중 감쇄 페널티에 의존하지 않는 용어 *w* 학습 과정에 영향을주지 않습니다. 가중치에 앞서 가우스를 사용한 MAP 베이지안 추론은 가중치 감소에 해당합니다.

전체 베이지안 추론과 마찬가지로 MAP 베이지안 추론은 이전에 가져온 정보를 활용하고 훈련 데이터에서 찾을 수없는 이점이 있습니다. 이 추가 정보는 MAP 포인트 추정치 (ML 추정치와 비교하여)의 분산을 줄이는 데 도움이됩니다. 그러나 편향이 증가하는 대가로 그렇게합니다.

가중치 감소로 정규화 된 최대 가능성 학습과 같은 많은 정규화 된 추정 전략은 베이지안 추론에 대한 MAP 근사치를 만드는 것으로 해석 될 수 있습니다. 이보기는 정규화가 다음에 해당하는 목적 함수에 추가 용어를 추가하는 것으로 구성된 경우에 적용됩니다. 로그 *피(θ).* 모든 정규화 페널티가 MAP 베이지안 추론에 해당하는 것은 아닙니다. 예를 들어, 일부 정규화 항은 확률 분포의 로그가 아닐 수 있습니다. 다른 정규화 용어는 데이터에 따라 달라지며, 물론 사전 확률 분포는 허용되지 않습니다.

MAP 베이지안 추론은 복잡하지만 해석 가능한 정규화 항을 설계하는 간단한 방법을 제공합니다. 예를 들어, 더 복잡한 페널티 항은 단일 가우시안 분포가 아닌 가우스 혼합을 사용하여 이전 (Nowlan과 Hinton , 1992 년).