PRML과 부수적인 자료들을 공부하여 간단히 정리하였습니다.

3.3 Bayesian Linear Regression

이전에 우리는 maximum likelihood 방법을 통해 linear regression 의parameter를 구하는 방법을 공부했다. 이는 몇 가지 특징(단점)이 있는데

- MLE 는 overfitting의 위험이 있다.
- 적절한 model complexity를 정해야 한다. by
  - basis function의 수
  - regularization coefficient
- 우리는 한정적인 dataset을 갖고 있기에 적절한 model complexity를 정하기 위해서는 cross validation과 같은 computationally expensive한 방법을 사용해야한다.

위와 같은 단점들을 해결하기 위해 우리는 Bayesian 방법론을 사용할 것이다.

- MAP는 uncertainty를 표현할 수 없기 때문에 distribution을 이용한다.

3.3.1 Parameter distribution

이전에 likelihood function $p(t/\mathbf{w})$ 이 Gaussian 이었다. 이에 대한 conjugate prior로 Gaussian을 가정한다. prior distribution은

$$p(\mathbf{w}) = N(\mathbf{m}_0, \mathbf{S}_0)$$

likelihood function과 prior를 곱해 posterior를 계산하면 (계산과정은 생략, monk영상을 보면 된다)

$$p(\mathbf{w}|\mathbf{t}) = N(\mathbf{m}_N, \mathbf{S}_N)$$

- $\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^T\mathbf{t})$
- $\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^T\mathbf{\Phi}$
- β : (target error term) noise precision parameter (assume known)

Gaussian은 mean과 mode(최빈값)가 같은 값을 갖기 때문에

- $\mathbf{w}_{MAP} = \mathbf{m}_N$ 이다.

만약 infinite broad prior인 경우 ($\mathbf{S}_0 = \alpha^{-1}\mathbf{I}, \alpha \to 0$) 수식을 전개해보면

- $\mathbf{m}_N \to \mathbf{m}_{ML}$

반대로 $N \to 0$ 이면 posterior 는 prior로 가까워진다.

복잡해 보이는 prior를 다소 간단한 형태로 정하면 $p(\mathbf{w}/\alpha) = N(0, \alpha^{-1}I)$ 으로 생각할 수 있다. 이 prior에서 posterior의 mean, precision은

$$\mathbf{m}_N = \beta\mathbf{S}_N\mathbf{\Phi}^T\mathbf{t}$$

$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}$$

log of posterior distribution (log of likelihood function과 log of prior의 합) 은

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + const$$

이는 결국 minimization of the sum of square with quadratic regulrarization($\lambda = \frac{\alpha}{\beta}$) 과 같은 수식이다.

3.3.2 Predictive distribution

우리의 최종 목표는 predictive distribution 이다.

$$p(t|\mathbf{t},\alpha,\beta) = \int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)d\mathbf{w}$$

predictive distribution을 보면 target의 conditional distribution $p(t/w,\beta) = N(t/y(w,x),\beta^{-1})$ 와 weight parameter $\mathbf{w}$의 posterior distribution으로 만들어졌다. 이를 토대로 정리하면

$$p(t|\mathbf{t},\alpha,\beta) = N(\mathbf{m}_N^T\phi(\mathbf{x}),\sigma_N^2(\mathbf{x}))$$

- variance : $\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \phi(\mathbf{x})^{\mathbf{T}}\mathbf{S_N}\phi(\mathbf{x})$

이 variance에서 첫번째 항은 data의 noise이고 (앞부분을 찾아보자) 뒷부분이 $\mathbf{w}$의 uncertainty를 나타낸다. noise와 $\mathbf{w}$ distribution은 independent하기에 두 값을 더한게 variance가 된것이다. N이 커질수록 posterior는 narrow해지므로 뒷부분은 0으로 간다.

3.3.3 Equivalent kernel

위에 w에 대해 구한 값을 토대로 mean of predictive distribution은

$$y(\mathbf{x},\mathbf{m}_N) = \mathbf{m}_N^T\phi(\mathbf{x}) = \beta\phi(\mathbf{x})^T\mathbf{S}_N\mathbf{\Phi}^T\mathbf{t} = \sum_{n=1}^{N}\beta\phi(\mathbf{x})^T\mathbf{S}_N\phi(\mathbf{x}_n)t_n$$

특정 $\mathbf{x}$에 대한 mean of predictive dist 은 결국 **training set target t의 linear combination** 이다. 이를 다르게 표현하면

$$y(\mathbf{x},\mathbf{m}_N) = \sum_{n=1}^{N}k(\mathbf{x},\mathbf{x}_n)t_n$$

- $k(\mathbf{x},\mathbf{x}') = \beta\phi(\mathbf{x})^T\mathbf{S}_N\phi(\mathbf{x}')$: 이 식을 *smoother matrix* or *equvalent kernel* 라고 부른다.
- 따라서 mean of predictive distribution at x 은 x와 (비슷한)가까운 data에 해당하는 t에 높은 가중치를 준다.

equvalent kernel에 대해서 covariance의 측면으로 살펴보자.

$$cov[y(\mathbf{x}),y(\mathbf{x}')] = cov[\phi(\mathbf{x})^T\mathbf{w},\mathbf{w}^T\phi(\mathbf{x}')] = \phi(\mathbf{x})^T\mathbf{S}_N\phi(\mathbf{x}') = \beta^{-1}k(\mathbf{x},\mathbf{x}')$$

equvalent kernel의 형태에서 근처의 points들의 predictive mean는 상관관계가 높다는 것을 알 수 있다.

3.4 Bayesian Model Comparison

Bayesian의 입장에서 model selection을 바라보고자 한다. 모델을 선택하는 과정에 있어서 확률적인 내용을 많이 사용한다.

- maximum likelihood와 관련한 overfitting 문제는 *marginalizing over the model parameters instead of making point estimates of their values* 로 해결 할 수 있다.
- 모델은 training data를 통해 바로 정할 수 있으므로 validation set이 필요없다. 따라서 모든 데이터를 학습시킬 수 있고 불필요한 검증과정을 없앨 수 있다.

이제 L개의 $\{M_i\}$ model이 있다고 하자. 이제 이 model들을 random variable으로 생각하고 model에 대한 uncertainty는 확률로 표현한다.

$$p(M_i|D) \propto p(M_i)p(D|M_i)$$

- 일단 model에 대한 prior는 다 같다고 가정하자.
- 따라서 우리의 주 관심은 **model evidence(=marginal likelihood)** : $p(D/M_i)$
  - model을 이루는 parameter들이 marginalized out 되었기에 marginal likelihood라고도 부름 (뒷 부분에 나옴)
- *Bayes factor*
  - ratio of model evidence s $p(D/M_i)p(D/M_j)$

model의 posterior를 알고 이를 이용하여 predictive distribution을 구하면

$$p(t|\mathbf{x}, D) = \sum_{i=1}^{L} p(t|\mathbf{x}, M_i, D)p(M_i|D)$$

이다. (mixture distribution의 모습) 이는 model posterior를 가중치로 하여 평균을 낸 것으로 볼 수 있다. 위와 같은 model averaging의 값과 가장 근사하는 좋은 model 하나를 찾고자 한다. 이를 *model selection* 이라고 한다.

- model evidence ($\mathbf{w}$는 model에 관한 parameter)
  - 이를 sampling 측면에서 바라보면, marginal likelihood는 data set D를 생성하는 probability로 볼 수 있는데 여기서 D는 prior로 부터 random하게 뽑힌 parameter들로 이루어진 model에서 만들어진 것이다.
  - 또한, evidence는 Bayes' Them에서 분모에 해당하는 normalizing term을 의미하기도 한다 : $p(\mathbf{w}/D, M_i) = \frac{p(D/\mathbf{w}, M_i)p(\mathbf{w}/M_i)}{p(D/M_i)}$

$$p(D|M_i) = \int p(D|\mathbf{w}, M_i)p(\mathbf{w}|M_i)d\mathbf{w}$$

이제 model evidence에 대해 더 알아보자.

- model이 single parameter w 를 갖고 있다고 가정
- notation M_i는 생략
- w의 posterior는 $p(D/w)p(w)$에 비례
- posterior는 w_{MAP} 에서 peaked 된 상태이고 그 때 width는 $\triangle\ w_{posterior}$ 라고 가정
- prior는 flat with width $\triangle\ w_{prior}$, 따라서 $p(w) = 1/\ \triangle\ w_{prior}$

$$p(D) = \int p(D|w)p(w)dw \simeq p(D|w_{MAP})\frac{1}{\triangle\ w_{prior}}\ \triangle\ w_{posterior}$$

log를 씌우면

$$\ln p(D) \simeq \ln p(D|w_{MAP}) + \ln(\frac{\triangle\ w_{posterior}}{\triangle\ w_{prior}})$$

- 첫번째 항 : 이 data를 가장 잘 표현하는 파라미터에 대한 확률값으로 log likelihood 의미
- 두번째 항 : model complexity에 대한 penalty 항
  - 우리는 $\ln p(D/M_i)$ 가 가장 큰 model(M_i)을 찾는 것이 목표이다. complex가 높은 model를 구하면 첫번째 항이 커질 것이지만 두번째 항은 $\triangle\ w_{posterior}$ 이 narrow 해지면서 음수가 되고 점점 작아진다. trade-off 관계인 것이다. 따라서 적절한 complexity가 있는 model을 선택하게 된다.
  - $\ln p(D|M_i) = accuracy(M_i) - complexity(M_i)$ 느낌
    - AIC, BIC를 예시로 생각할 수 있다.
- M 개의 parameter가 있을 경우
  - 위에서 설명한 부분과 같다. 뒷 부분에 M이 추가되어 M이 커지면서 더 penalty를 준다.

$$\ln p(D/\mathbf{w}_{MAP}) + M\ln(\frac{\triangle\ w_{posterior}}{\triangle\ w_{prior}})$$

optimal model complexity (model selection) 는 maximum evidence 으로 정해진다는 것 을 기억하자.

3.5 The Evidence Approximation

linear basis model에서 완전한 Bayesian 접근법을 위해서 $\mathbf{w}$에 대한 hyperparameter α, β의 prior를 고려해보자.

- predictive distribution은 아래와 같이 구할 수 있다. ($\mathbf{x}$ 표시는 생략)
  - $p(t/\mathbf{w})$: distribution of target ($= N(y(x, \mathbf{w}), \beta^{-1})$)
  - $p(\mathbf{w}/\mathbf{t}, \alpha, \beta)$: posterior of $\mathbf{w}$
  - $p(\alpha, \beta/\mathbf{t})$: posterior of α, β

- - -

$$p(t|\mathbf{t}) = \iiint p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)p(\alpha,\beta|\mathbf{t})d\mathbf{w}d\alpha d\beta$$

하지만 여기서 문제가 발생한다. 위처럼 모든 변수에 대해 marginalize하는 것은 항상 가능한 것이 아니다 (analytically intractable). 그래서 우리는 hyperparameter를 특정한 값으로 approximation한다. 그 방법은 maximizing marginal likelihood function이다. 이러한 방법론을 *evidence approximation* (통계에서는 *emprical Bayes, type 2 maximum likelihood, generalized maximum likelihood*) 라고 부른다.

만약에 posterior distribution $p(\alpha,\beta/\mathbf{t})$ 가 특정한 값 $\hat{\alpha},\hat{\beta}$에서 가장 높은 값(peaked)을 가진다면 predictive distribution은 $\mathbf{w}$에 대해서만 marginalize해서 구할 수 있을 것이다.

$$p(t|\mathbf{t}) \simeq p(t|\mathbf{t},\hat{\alpha},\hat{\beta}) = \int p(t|\mathbf{w},\hat{\beta})p(\mathbf{w}|\mathbf{t},\hat{\alpha},\hat{\beta})d\mathbf{w}$$

posterior distribution for α,β 는

$$p(\alpha,\beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha,\beta)p(\alpha,\beta)$$

prior는 flat 하다고 가정하자. 따라서 우리는 $\hat{\alpha},\hat{\beta}$를 구하기 위해서 marginal likelihood function $p(\mathbf{t}/\alpha,\beta)$ 을 최대로 만드는 찾으면 된다. 이를 통해 우리는 cross validation과 같은 방법이 아니라 한 번에 hyperparameter를 찾을 수 있다. 찾는 방법은 미분을 이용하는 방법, EM 알고리즘을 이용하는 방법이 있다. 전자는 이제 살펴볼 것이고 후자는 9장에서 배운다.

3.5.1 Evaluation of the evidence function

marginal likelihood function은 parameter $\mathbf{w}$를 marginalize해서 얻을 수 있다.

- $p(\mathbf{t}/\mathbf{w},\beta)$: likelihood function
- $p(\mathbf{w}/\alpha)$: prior of w

$$p(\mathbf{t}|\alpha,\beta) = \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)d\mathbf{w}$$

위의 식을 Gaussian의 형태를 이용하여 정리해보자.

$$p(\mathbf{t}|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\{-E(\mathbf{w})\}d\mathbf{w}$$

$$E(\mathbf{w}) = \beta E_D(\mathbf{w}) + \alpha E_w(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t}-\mathbf{\Phi w}\|^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^T\mathbf{A}(\mathbf{w}-\mathbf{m}_N)$$

- $\mathbf{A} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}$
- $\mathbf{m}_N = \beta\mathbf{A}^{-1}\mathbf{\Phi}^T\mathbf{t}$
  - 이전에 구한 식과 동일하다 (3.3 참고)

이제 이를 이용하면

$$\begin{aligned}\int \exp\left(-E(\mathbf{w})\right)d\mathbf{w} &= \exp(-E(\mathbf{m}_N))\int \exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^T\mathbf{A}(\mathbf{w}-\mathbf{m}_N)\right\}d\mathbf{w} \\ &= \exp\left\{-E(\mathbf{m}_N)\right\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2}\end{aligned}$$

이를 이용하여 최종적으로 log marginal likelihood function을 구하면

$$\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_n) - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln(2\pi)$$

3.5.2 Maximizing the evidence function

$\ln p(\mathbf{t}/\alpha,\beta)$을 최대화하는 α,β를 구하기 위해 미분을 이용해보자.

- $(\beta\Phi^T\Phi)\mu_i = \lambda_i\mu_i$ 라고 하면 $\mathbf{A}$의 eigenvalue는 $\alpha + \lambda_i$ 이다. 따라서

$$\frac{d}{d\alpha}\ln|\mathbf{A}| = \frac{d}{d\alpha}\ln\prod_i(\lambda_i+\alpha) = \frac{d}{d\alpha}\sum_i\ln(\lambda_i+\alpha) = \sum_i\frac{1}{\lambda_i+\alpha}$$

- α에 대해 미분

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\mathbf{m}_N^T\mathbf{m}_N - \frac{1}{2}\sum_i\frac{1}{\lambda_i+\alpha}$$

$$\alpha\mathbf{m}_N^T\mathbf{m}_N = M - \alpha\sum_{i=1}^{M}\frac{1}{\lambda_i+\alpha} = \gamma$$

이를 다시 정리하면

$$\gamma = \sum_{i=1}^{M}\frac{\lambda_i}{\alpha+\lambda_i}$$

최종적으로 α에 대해 정리하면

$$\alpha = \frac{\gamma}{\mathbf{m}_N^T\mathbf{m}_N}$$

그런데 $\gamma, \mathbf{m}_N$ 모두 α에 depend한다. 따라서 이를 위해서 iterative한 방법을 사용한다. 임의의 수로 α를 시작하고 $\gamma, \mathbf{m}_N$을 구한다. 다시 이 두 값으로 α를 구한다. 이렇게 수렴할 때까지 반복하는 것이다.

- β에 대해 미분

$$\frac{d}{d\beta}\ln|\mathbf{A}| = \frac{d}{d\beta}\sum_i\ln(\lambda_i+\alpha) = \frac{1}{\beta}\sum_i\frac{\lambda_i}{\lambda_i+\alpha} = \frac{\gamma}{\beta}$$

$$0 = \frac{N}{2\beta} - \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{m}_N^T\phi(\mathbf{x}_n)\}^2 - \frac{\gamma}{2\beta}$$

$$\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N}\{t_n - \mathbf{m}_N^T\phi(\mathbf{x}_N)\}^2$$

이도 마찬가지도 iterative하게 구한다.

- cross validation과 같은 **추가적인 computation이 없이 한 번에 train data set을 모두 이용하여 model complexity를 정할 수 있다** 는 점을 기억하자.

3.5.3 Effective number of parameters

- 이 부분은 ridge regression의 내용 (data의 고윳값이 작은 방향의 parameter가 0에 가까워진다) 과 같다. ESL 책의 linear regression부분을 보면 알 수 있다.

α에 대한 Bayesian의 접근에 대해 조금 더 살펴보자. $\beta\Phi^T\Phi$는 positive definite matrix이므로 eigenvalue가 모두 0이상의 값을 갖는다. 따라서

- $0 \le \lambda_i/(\lambda_i+\alpha) \le 1$
- $0 \le \gamma \le 1$

임을 알 수 있다. $\lambda_i >> \alpha$ 인 경우는 이에 해당하는 parameter w_i가 maximum likelihood의 값과 가까워지고 $\lambda_i/(\lambda_i+\alpha)$ 이 1에 가까워진다. 반대의 경우는 $w_i, \lambda_i/(\lambda_i+\alpha)$ 모두 0에 가까워진다.

- 따라서 γ는 measures the effective total number of well determined parameters

다음은 β에 대해 알아보자. 위에서 보았듯이 effective number of parameter는 γ이고 나머지 $M-\gamma$개의 parameters 들이 prior에 의해 작은 값을 갖는다. 이것이 variance에서 $\frac{1}{N-\gamma}$로 나타나고 bias of maximum likelihood result를 바로 잡아준다.

- 만약에 $N >> M$의 상황인 경우, 대부분의 parameter들이 well determined될 것이고 data size에 따라 eigenvalue도 커지게 된다. 그러면 $\gamma = M$이 되고 evidence approximation도 아래 값을 이용해 간단해진다. (data 많은게 짱이다)

$$\alpha = \frac{M}{2E_W(\mathbf{m}_N)}$$

$$\beta = \frac{N}{2E_D(\mathbf{m}_N)}$$

3.6 Limitations of Fixed Basis Functions

- 장점
  - nonlinear basis functions의 linear combination이니까 해석이 쉽다.
  - closed form의 해가 존재한다.
- 단점
  - basis function이 training data를 보기 전에 이미 fixed되서 시작한다.
  - 차원의 저주
  - input간의 correlation 때문에 보다 작은 차원에 nonlinear manifold에 데이터가 분포할 수 있다.