LECTURE 17 - BAYESIAN REGRESSION & ML, MAP CONTINUED...

1. Regression, cont.

- Look back our polynomial regression:

$$\min E^*(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(y(x_n,\mathbf{w}) - t_n\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \tag{1}$$

  This is equivalent to:

$$\max \prod_{n=1}^{N} \exp\left\{-\frac{1}{2}\left(y(x_n,\mathbf{w}) - t_n\right)^2\right\} \exp\left\{-\frac{\lambda}{2}\|\mathbf{w}\|_2^2\right\} \tag{2}$$

- As discussed, the first term is the Likelihood and the second term is the prior on the weights
- These are Gaussian distributions:

$$\mathscr{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} \tag{3}$$

- σ^2 is the variance OR $\frac{1}{\sigma^2}$ is the *precision*
- So, as λ gets big, variance gets smaller/tighter. As λ gets small, variance gets larger/wider.
- Previously, we used:

$$y = \sum_{j=0}^{M} w_j x^j \tag{4}$$

- We can extend this, make is more general and flexible:

$$y = \sum_{j=0}^{M} w_j \phi_j(\mathbf{x}) \tag{5}$$

  where $\phi_j(\mathbf{x})$ is a *basis function*
- For example:
  - Basis function we were using previously: $\phi_j(x) = x^j$ (for univariate x)
  - Linear Basis Function: $\phi_j(\mathbf{x}) = x_j$
  - Radial Basis Function: $\phi_j(\mathbf{x}) = \exp\left\{-\frac{(x-\mu_j)^2}{2s_j^2}\right\}$
  - Sigmoidal Basis Function: $\phi_j(\mathbf{x}) = \frac{1}{1+\exp\left\{\frac{\mathbf{x}-\mu_j}{s}\right\}}$

- As before:

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \tag{6}$$

- However, now:

$$y = \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}) = [w_0, w_1, \ldots, w_M][\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x})]^T \tag{7}$$

where $\epsilon \sim \mathscr{N}(\cdot|0, \beta^{-1})$

$$p(t|\mathbf{w}, \beta) = \prod_{n=1}^{N} \mathscr{N}(t_n|\mathbf{w}^T \mathbf{\Phi}(\mathbf{x}_n), \beta^{-1}) \tag{8}$$

- So, what is the "trick" to use to maximize this?

$$\mathscr{L} = \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E(\mathbf{w}) \tag{9}$$

$$\frac{\partial \mathscr{L}}{\partial \mathbf{w}} = \beta \sum_{n=1}^{N} (t_n - \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}_n)) \mathbf{\Phi}(\mathbf{x}_n)^T = 0 \tag{10}$$

- This results in:

$$\mathbf{w}_{ML} = \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{t} \tag{11}$$

where

$$\mathbf{\Phi} = [\mathbf{\Phi}(x_1), \mathbf{\Phi}(x_2), ...] \tag{12}$$

- What would you do if you want to include a prior? get the MAP solution? If assuming zero-mean Gaussian noise, then Regularized Least Squares!

2. Bayesian Linear Regression

- Recall: $E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$ where λ is the trade-off regularization parameter
- A simple regularizer (and the one we used previously) is: $E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$
- If we assume zero-mean Gaussian noise: $E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \left\{t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n)\right\}^2$
- Then, the total error becomes: $\frac{1}{2}\sum_{n=1}^{N} \left\{t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n)\right\}^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$
- We can take the derivative, set it equal to zero and solve for the weights. When we do, we get:

$$\mathbf{w} = \left(\lambda \mathbf{I} + \mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{t} \tag{13}$$

- Recall, we can interpret this as:

$$\begin{aligned}
\min_{\mathbf{w}} E^* &= \min_{\mathbf{w}} \{E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})\} && (14)\\
&= \max_{\mathbf{w}} \{-E_D(\mathbf{w}) - \lambda E_W(\mathbf{w})\} && (15)\\
&= \max_{\mathbf{w}} \exp\{-E_D(\mathbf{w}) - \lambda E_W(\mathbf{w})\} && (16)\\
&= \max_{\mathbf{w}} \exp\{-E_D(\mathbf{w})\} \exp\{-\lambda E_W(\mathbf{w})\} && (17)\\
&\propto \max_{\mathbf{w}} \prod_{n=1}^{N} \mathcal{N}\left(t \left| \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}_n), \beta \mathbf{I}\right.\right) \mathcal{N}\left(\mathbf{w} \left| \mathbf{m}_0, \mathbf{S}_0\right.\right) && (18)\\
&= \max_{\mathbf{w}} p\left(\mathbf{t} \left| \mathbf{w}, \mathbf{X}\right.\right) p\left(\mathbf{w}\right) && (19)\\
&\propto \max_{\mathbf{w}} p(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) && (20)
\end{aligned}$$

  where $\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0 \mathbf{m}_0 + \beta \mathbf{\Phi}^T \mathbf{t}\right)$ and $\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta \mathbf{\Phi}^T \mathbf{\Phi}$
- What happens with different values of β and $\mathbf{S}_0$?
- To simplify, let us assume that $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$ and $\mathbf{m}_0 = \mathbf{0}$, thus, $\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^T \mathbf{t}$ and $\mathbf{S}_N^{-1} = (\alpha^{-1}\mathbf{I})^{-1} + \beta \mathbf{\Phi}^T \mathbf{\Phi} = \alpha \mathbf{I} + \beta \mathbf{\Phi}^T \mathbf{\Phi}$
- This results in the following Log Posterior:

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}_n)\right)^2 - \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} + const \qquad (21)$$

- Let us suppose we are dealing with 1-D data, $\mathbf{X} = \{x_1, \ldots, x_N\}$ and a linear form for y: $y(x, \mathbf{w}) = w_0 + w_1 x$
- We are going to generate synthetic data from: $t = -0.3 + 0.5x + \epsilon$ where ϵ is from zero-mean Gaussian noise. The goal is to estimate the true values $w_0 = -0.3$ and $w_1 = 0.5$.
- *Let us assume $\beta = 25$ and $\alpha = 2$ and run provided code example and step through it carefully.*