Tutorial 2

CS 337 Artificial Intelligence & Machine Learning, Autumn 2019

Week 2, August, 2019

Problem 1. Consider a data set in which each data point y_i is associated with a weighting factor r_i, so that the sum-square error function becomes

$$\frac{1}{2}\sum_{i=1}^{m} r_i(y_i - w^T\phi(x_i))^2$$

Find an expression for the solution w^* that minimizes this error function. The weights r_i's are known before hand. (Exercise 3.3 of Pattern Recognition and Machine Learning, Christopher Bishop).

Solution:

Let $r_{m+1} = 1$ and let R be an $(m+1) \times (m+1)$ diagonal matrix of $r_1, r_2, \ldots, r_{m+1}$.

$$R = \begin{bmatrix} r_1 & 0 & ... & 0 & \\ 0 & r_2 & ... & 0 & \\ ... & ... & ... & ... & 1 \\ 0 & 0 & 0 & ... & r_{m+1} \end{bmatrix}$$

Further, let

$$\Phi = \begin{bmatrix} \phi_1(x_1) & ... & \phi_p(x_1) & 1 \\ ... & ... & ... & 1 \\ \phi_1(x_m) & ... & \phi_p(x_m) & 1 \end{bmatrix}$$

and

$$\widehat{\mathbf{w}} = \begin{bmatrix} w_1 \\ ... \\ w_p \\ b \end{bmatrix}$$

and

$$\mathbf{y} = \begin{bmatrix} y_1 \\ ... \\ y_m \end{bmatrix}$$

The sum-square error function then becomes

$$\frac{1}{2}\sum_{i=1}^{m} r_i(y_i - (\widehat{\mathbf{w}}^T\phi(x_i) + b))^2 = \frac{1}{2}||\sqrt{R}\mathbf{y} - \sqrt{R}\Phi\widehat{\mathbf{w}}||_2^2$$

where $\sqrt{R}$ is a diagonal matrix such that each diagonal element of $\sqrt{R}$ is the square root of the corresponding element of R. This is a convex function being minimized (prove this using techniques similar to what we employed for least squares linear regression) and therefore has a global minimum at $\widehat{\mathbf{w}}_*^{x'}$ where the gradient must become 0. (again work out the steps using techniques similar to what we employed for least squares linear regression). The expression for the solution $\widehat{\mathbf{w}}^*$ that minimizes this error function is therefore

$$\widehat{\mathbf{w}}_*^{x'} = (\Phi^T R\Phi)^{-1}\Phi^T R\mathbf{y}$$

Problem 2. Equivalence between Ridge Regression and Bayesian Linear Regression (with fixed σ^2 and λ): Consider the Bayesian Linear Regression Model

$$\begin{aligned} y = \mathbf{w}^T\phi(\mathbf{x}) + \varepsilon \text{ and } \varepsilon \sim \mathcal{N}(0, \sigma^2) \\ \mathbf{w} \sim \mathcal{N}(0, \alpha I) \text{ and } \mathbf{w} \mid \mathcal{D} \sim \mathcal{N}(\mu_m, \Sigma_m) \\ \mu_m = (\lambda\sigma^2 I + \phi^T\phi)^{-1}\phi^T\mathbf{y} \text{ and } \Sigma_m^{-1} = \lambda I + \phi^T\phi/\sigma^2 \end{aligned}$$

Show that $\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\text{argmax}}\ \Pr(\mathbf{w} \mid \mathcal{D})$ is the same as that of *Regularized Ridge Regression.*

$$\mathbf{w}_{Ridge} = \underset{\mathbf{w}}{\text{argmin}}\ ||\phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda\sigma^2||\mathbf{w}||_2^2$$

In other words, The Bayes and MAP estimates for Linear Regression coincide with that of *Regularized Ridge Regression.*

Solution Sketch: Taking the negative log of the log likelihood we see that maximizing the log of the posterior distribution is equivalent to minimizing the ridge regression objective.

$$\Pr(\mathbf{w} \mid \mathcal{D}) = \mathcal{N}(\mathbf{w} \mid \mu_m, \Sigma_m) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_m|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{w}-\mu_m)^T\Sigma_m^{-1}(\mathbf{w}-\mu_m)}$$

$$-\log\Pr(\mathbf{w}) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma_m| + \frac{1}{2}(\mathbf{w} - \mu_m)^T\Sigma_m^{-1}(\mathbf{w} - \mu_m)$$

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\text{argmax}} -\log\Pr(\mathbf{w}) = \underset{\mathbf{w}}{\text{argmax}}\frac{1}{2}\mathbf{w}^T\Sigma_m^{-1}\mathbf{w} - \mathbf{w}^T\Sigma_m^{-1}\mu_m$$

that is,

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\text{argmax}}\frac{1}{2}\mathbf{w}^T\left(\lambda I + \phi^T\phi/\sigma^2\right)\mathbf{w} - \mathbf{w}^T\left(\lambda I + \phi^T\phi/\sigma^2\right)\left((\lambda\sigma^2 I + \phi^T\phi)^{-1}\phi^T\mathbf{y}\right)$$

and after expanding and canceling out redundant terms, and later, after completing squares:

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\text{argmax}} \frac{1}{2\sigma^2} \mathbf{w}^T \left(\phi^T \phi \mathbf{w} - 2\phi^T \mathbf{y}\right) + \lambda \mathbf{w}^T \mathbf{w} = \underset{\mathbf{w}}{\text{argmax}} \frac{1}{2} ||\phi \mathbf{w} - \mathbf{y}||^2 + \sigma^2 \lambda ||\mathbf{w}||^2 = \mathbf{w}_{Ridge}$$

Problem 3. Ridge Regression and Error Minimization:

1. *Prove the following Claim:*
   The sum of squares error on training data using the weights obtained after minimizing ridge regression objective is greater than or equal to the sum of squares error on training data using the weights obtained after minimizing the ordinary least squares (OLS) objective.

   More specifically, if ϕ and $\mathbf{y}$ are defined on the training set $\mathcal{D} = \{(\mathbf{x}_1, y_1)...(\mathbf{x}_m, y_m)\}$ as

$$\phi = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & ..... & \phi_n(\mathbf{x_1}) \\ . & & & \\ . & & & \\ \phi_1(\mathbf{x}_m) & \phi_2(\mathbf{x}_m) & ..... & \phi_n(\mathbf{x}_m) \end{bmatrix} \tag{1}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ . \\ y_m \end{bmatrix} \tag{2}$$

   and if

$$\mathbf{w}_{Ridge} = \underset{\mathbf{w}}{\text{argmin}} \ ||\phi \mathbf{w} - \mathbf{y}||_2^2 + \lambda ||\mathbf{w}||_2^2$$

   and

$$\mathbf{w}_{OLS} = \underset{\mathbf{w}}{\text{argmin}} \ ||\phi \mathbf{w} - \mathbf{y}||_2^2$$

   then you should prove that

$$||\phi \mathbf{w}_{Ridge} - \mathbf{y}||_2^2 \geq ||\phi \mathbf{w}_{OLS} - \mathbf{y}||_2^2$$

   Solution: If

$$\mathbf{w}_{OLS} = \underset{\mathbf{w}}{\text{argmin}} \ ||\phi \mathbf{w} - \mathbf{y}||_2^2$$

   then by definition of argmin,

$$||\phi \mathbf{w}_{Ridge} - \mathbf{y}||_2^2 \geq ||\phi \mathbf{w}_{OLS} - \mathbf{y}||_2^2$$

   Also, one can reformulate

$$\mathbf{w}_{Ridge} = \underset{\mathbf{w}}{\text{argmin}} \ ||\phi \mathbf{w} - \mathbf{y}||_2^2 + \lambda ||\mathbf{w}||_2^2$$

as

$$\mathbf{w}_{Ridge} = \underset{\mathbf{w}}{\operatorname{argmin}} \ ||\phi\mathbf{w} - \mathbf{y}||_2^2$$

$$such\ that\ ||\mathbf{w}||_2^2 \leq \theta$$

for some θ corresponding to a value of λ. The solution to a constrained minimization problem will always be greater than or equal to its unconstrained counterpart.

2. If it is the case that ridge regression leads to greater error than ordinary least squares regression, then why should one be interested in ridge regression at all?

   Solution: This is still acceptable since ridge regression incorporates prior (as per Bayesian interpretation). The idea is ultimately to do well on unseen (test) data. Therefore, higher training error might be acceptable if test error can be lowered.

Problem 4. Gradient descent is a very helpful algorithm. But it is not guaranteed to converge to global minima always. Give an example of a continuous function and initial point for which gradient descent converges to a value which is not global minima.

Problem 5. In class, we have illustrated Bayesian estimation for the parameter μ of a Normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, assuming that σ was known by imposing a Normal (conjugate) prior on μ. Now suppose that the parameter μ is known and we wish to estimate σ^2. What will be the form of the conjugate prior for this estimation procedure? If $\mathcal{D} = X_1, X_2, X_3, ..., X_n$ is a set of independent samples from this distribution, after imposing the conjugate prior, compute the form of the likelihood function $\mathcal{L}(\theta)$, the posterior density $P(\theta \mid \mathcal{D})$ and the posterior probability $P(X \mid \mathcal{D})$. Again, you can ignore normalization factors.

Solution:

- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 ... x_m$
- $\mu_{MLE} = \frac{1}{m} \sum_{i=1}^{m} x_i$ and $\sigma^2_{MLE} = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_{MLE})^2$
- Suppose you are told that σ^2 is a random variable is and μ is not.

$$\Pr(x_i|\mu;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\Pr(\mathcal{D}|\mu) = \left(\frac{1}{(2\pi)^{\frac{m}{2}}(\sigma^{\mathbf{2}})^m}\right) exp\left(\frac{-1}{2\sigma^{\mathbf{2}}}\sum_{i=1}^{m}(x_i-\mu)^2\right)$$

Note the positions of $\sigma^{\mathbf{2}}$ in the likelihood above. In order to make the posterior of the same form as the prior, we should make the prior jell seamlessly with the likelihood because $\Pr(\theta|\mathcal{D}) \propto \Pr(\mathcal{D}|\theta)\Pr(\theta)$, where we have $\theta := \sigma^{\mathbf{2}}$. This could mean

$$\Pr(\theta) \propto \frac{1}{\theta^A} exp(\frac{-B}{\theta})$$

One can normalize this distribution to find the proportionality constant.

It is ok if the students get so far in suggesting the prior. It is also ok if the students miss somehow land up only with $\Pr(\theta) \propto \frac{1}{\theta} exp(\frac{-B}{\theta}))$

Part 2 of the question:

$$\Pr(x|D) = \int_{\theta} \Pr(x|\theta)\Pr(\theta|D)d\theta$$

Substituting,

$$\Pr(x|D) = \int_{\sigma^{\mathbf{2}}} \Pr(x|\sigma^{\mathbf{2}})\Pr(\sigma^{\mathbf{2}}|D)d\sigma^{\mathbf{2}} = \int_{\sigma^{\mathbf{2}}} \Pr(x|\sigma^{\mathbf{2}})\Pr(\sigma^{\mathbf{2}}|D)d\sigma^{\mathbf{2}}$$

We can substitute and leave the integral as it is. An approximation is to use the MAP or Bayes estimate in place of integration and

$\Pr(x|D) \approx \Pr(x|\sigma^{\mathbf{2}}_{\mathbf{MAP}})\Pr(\sigma^{\mathbf{2}}_{\mathbf{MAP}}|D)$

No need to give marks to what follows: This is called an inverse-gamma distribution.

$$p(\theta) = \frac{B^{A-1}}{\Gamma(A-1)}\frac{1}{\theta^A} exp(\frac{-B}{\theta})$$

The posterior is also an inverse-gamma distribution with

$$A' = A + n/2$$

$$B' = B + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Marginal likelihood

$$p(\mathcal{D}) = \frac{\frac{1}{\sqrt[2]{2\pi\theta}} exp(\sum_{i=1}^{n}\frac{-(x_i-\mu)^2}{2\theta})\frac{B^{A-1}}{\Gamma(A-1)}\frac{1}{\theta^A}exp(\frac{-B}{\theta})}{\frac{B'^{A'-1}}{\Gamma(A'-1)}\frac{1}{\theta^{A'}}exp(\frac{-B'}{\theta})}$$

Posterior Predictive

$$p(x|\mathcal{D}) = \frac{p(x,\mathcal{D})}{p(\mathcal{D})}$$

Use $\mathcal{D}' = (\mathcal{D}, x)$ to compute $p(\mathcal{D}')$ and substitute back to get

$$p(x|\mathcal{D}) = t_{2A'}(x|\mu, \theta = \frac{B'}{A'})$$

Problem 6. Consider a linear model of the form

$$y(x, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$$

together with a sum-of-squares error function of the form

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$ (i.e. $\mathbb{E}[\epsilon_i \epsilon_j] = \sigma^2$ when $i = j$), show that minimizing E_D averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter w_0 is omitted from the regularizer. (Problem 3.4 from Bishop, PRML)

Solution:
After adding Gaussian noise to each of the input variables, let:

$$\bar{y}_n = w_0 + \sum_{i=1}^{D} w_i (x_{ni} + \epsilon_{ni}) = y_n + \sum_{i=1}^{D} w_i \epsilon_{ni}$$

where $y_n = y(x_n, \mathbf{w})$ and $\epsilon_{ni} \sim \mathcal{N}(0, \sigma^2)$.
The sum-of-squares error function becomes:

$$\begin{aligned} \bar{E} &= \frac{1}{2} \sum_{n=1}^{N} \{\bar{y}_n - t_n\}^2 \\ &= \frac{1}{2} \sum_{n=1}^{N} \left\{ y_n^2 + 2y_n \sum_{i=1}^{D} w_i \epsilon_{ni} + \left(\sum_{i=1}^{D} w_i \epsilon_{ni} \right)^2 - 2t_n y_n - 2t_n \sum_{i=1}^{D} w_i \epsilon_{ni} + t_n^2 \right\} \end{aligned} \tag{3}$$

Taking the expectation of $\bar{E}$ under ϵ_{ni}, the second and fifth terms in Equation 3 disappear (because $\mathbb{E}[\epsilon_{ni}] = 0$) and for the third term we get $\mathbb{E}\left[\left(\sum_{i=1}^{D} w_i \epsilon_{ni} \right)^2 \right] = \sum_{i=1}^{D} w_i^2 \sigma^2$ (because ϵ_{ni} are all independent with variance σ^2). Thus we have:

$$\mathbb{E}[\bar{E}] = E_D + \frac{1}{2} \sum_{i=1}^{D} w_i^2 \sigma^2$$

which is what was asked for.