Exercise 1.1: Maximum Likelihood and Overfitting (*14 points*)

This exercise is intended to help you obtain an intuitive understanding of overfitting and its consequences. Consider the following polynomial model of order P:

$$y_n = \theta_0 + \theta_1 x_n + \theta_2 x_n^2 + \cdots + \theta_P x_n^P + \varepsilon_n, \ n = 1, \ldots, N \tag{1}$$

where ε_n denotes an iid. Gaussian noise with zero mean and variance σ^2.

(a) Derive the log-likelihood function for the model from eq (1). *Hint*: Express the measurement noise in terms of the data and the model parameters and plug the expression into the noise probability density. (*2 points*)

(b) Using the log-likelihood function from (a), derive the maximum likelihood (ML) estimate for the model parameters θ_p from eq (1). *Hint*: The log-likelihood from (a) is a function of the parameters θ_0 to θ_P. Derive an expression for the values of θ_0 to θ_P that maximize the log-likelihood. You may find the list of matrix derivatives in [1] useful. (*3 points*)

The remaining part of this exercise require a computer. Next consider the following quadratic model:

$$y_n = \theta_0 + \theta_1 x_n + \theta_2 x_n^2 + \varepsilon_n. \tag{2}$$

For now, we will fix the model parameters to the following values: $\theta_0 = 0.3$, $\theta_1 = -0.1$, $\theta_2 = 0.5$ and $\sigma^2 = 0.001$. Let x go from -0.5 to 0.2 in steps of 0.1 (i.e.: $x_1 = -0.5, x_2 = -0.4, \ldots$).

(c) Generate data from the model in eq (2) using the values for x, θ_0 to θ_2 and σ^2 given above. *Hint*: Use a random number generator (like the `randn` command in MATLAB) to generate normally distributed random values for the noise term ε_n. (*1 point*)

(d) Take the ML estimator for the P^{th}-order model derived in (b), set P to 1, 2 and 7, and apply it to the data from (c). What do you notice when comparing the ML parameter estimates to their true values? Calculate the log-likelihood for the data from (c) and the ML parameter estimates from above. What do you notice about the relationship between the log-likelihood and P? *Note*: If you could not solve (b), you may use a general purpose optimizer like MATLAB's `fminsearch` to find the ML solution. (*3 points*)

Now, increase x from -0.5 to 0.5 in steps of 0.01, but keep the values of θ_0 to θ_2 and σ^2 the same as above.

(e) Generate data from the model in eq (2) using the new values of x, but do not apply the ML estimator to the new data. Instead, calculate the log-likelihood for the new data under the old ML parameter estimates obtained in (d). What do you observe now about the relationship between the log-likelihood and P? (*2 points*)

(f) Now, apply the ML estimator to the new data and compare the new ML estimates to the old estimates from (d) and the true parameters. Also calculate the log-likelihood for the new data under the new ML estimates. What do you observe about the relationship between the log-likelihood and P? (*3 points*)

Exercise 1.2: Maximum-A-Posteriori Estimation (*11 points*)

This exercise illustrates the regularizing effect of placing a prior distribution over model parameters. Consider the polynomial model from exercise 1.1 (eq (1)). Let's collect the model parameters into a vector:

$$\boldsymbol{\theta} = (\theta_0, \ldots, \theta_p, \ldots, \theta_P)^T \,.$$

In this vector notation, a Gaussian prior over the parameters is given by:

$$p(\boldsymbol{\theta}) = \frac{1}{\sqrt{|2\pi\Sigma_0|}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^T \Sigma_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0)\right), \tag{3}$$

where Σ_0 is the prior covariance and $\boldsymbol{\mu}_0$ the prior mean. For the purpose of this exercise, we will set $\Sigma_0 = I$ and $\boldsymbol{\mu}_0 = 0$, where I denotes the identity matrix. This leads to so called shrinkage priors.

(a) Write down the log-posterior distribution $\log p(\boldsymbol{\theta}|y)$ for the model from eq (1) with the prior from eq (3). *Hint*: Start with Bayes rule. Do not evaluate the model evidence $p(y)$. (*3 points*)

(b) Using the log-posterior distribution from (a), derive the maximum-a-posteriori (MAP) estimate for the model parameters $\boldsymbol{\theta}$ from eq (1) with the prior from eq (3). *Hint*: You may find the list of matrix derivatives in [1] useful. (*3 points*)

The following part requires a computer.

(c) Apply the MAP estimator derived in (b) to the data from exercise 1.1 (c) for $P = 1, 2$ and 7. What do you notice when comparing the MAP estimates to the ML estimates from exercise 1.1 (d) and the true values of the parameters? *Note*: If you could not solve (b), you may use a general purpose optimizer to find the MAP estimate. (*3 points*)

(d) Apply the MAP estimator to the data from exercise 1.1 (e) and compare the MAP estimates to the ML estimates. What do you notice about the difference between ML and MAP estimates for the data from 1.1 (c) and (e)? (*2 points*)

Exercise 1.3: Bayesian Inference in the Univariate Gaussian Case (*10 points*)

In this exercise, you will use Bayesian inference to analytically invert a simple model. Consider the following univariate Gaussian model:

$$y_n = x\theta + \varepsilon_n, \; n = 1, \ldots, N. \tag{4}$$

Note that x is a constant scaling factor independent of n. Noise and prior distributions are Gaussian:

$$p(\varepsilon_n) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{\varepsilon_n^2}{2\sigma_\varepsilon^2}\right) \tag{5}$$

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(\theta - \mu_p)^2}{2\sigma_p^2}\right) \tag{6}$$

(a) Derive the likelihood for this model. *Hint*: Combine eqs (4) and (5). (*2 points*)

(b) Using prior and likelihood, derive an expression for the log-posterior distribution $\log p(\theta|y_1, \ldots, y_N)$. *Hint*: Start with Bayes rule. Do not evaluate the model evidence $p(y_1, \ldots, y_N)$. Writing the sum of all observations as N times the mean will simplify the expression ($\sum_{n=1}^{N} y_n = N\bar{y}$). (*3 points*)

(c) Compare the expression from (b) with the log-distribution of a standard Gaussian $\log N(\theta|\mu, \sigma^2)$. What do you notice about the dependence on θ? *Hint*: Eq (6) defines a so called conjugate prior to the likelihood you derived in (a). (*2 points*)

(d) Derive expressions for the parameters μ and σ^2 of the posterior distribution by comparing the coefficients for the first and second powers of θ in the standard Gaussian to those in the log-posterior from (b). *Hint*: This procedure is known as "completing the square". (*3 points*)

References

[1] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," 2012. [Online]. Available: http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf