N. Bertschinger
M. Kaschube
V. Ramesh

Machine Learning I
4. Exercise Sheet

Exercise 1. *Show that the Dirichlet distribution supported on* $\mathbf{p} = (p_1, \ldots, p_K) \subseteq (0,1)^K$ *with* $\sum_{i=1}^K p_i = 1$ *and probability density function*

$$p(\mathbf{p}; \alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K p_i^{\alpha_i - 1}$$

is the conjugate prior for the categorical distribution on K *outcomes:*

$$p(X = i \mid \mathbf{p}) = p_i$$

Why are the parameters $\alpha = (\alpha_1, \ldots, \alpha_K)$ *often called* pseudo-counts*?*
Hint: *Consider that the likelihood of a data set* $D = \{x_n\}_{n=1}^N$*, where each* $x_n \in \{1, \ldots, K\}$*, can be written as*

$$p(D) = \prod_{k=1}^K p_k^{\#\{x_n = k \,:\, x_n \in D\}}$$

2 points

Exercise 2. *Implement Ridge regression and illustrate the effect of the regularization parameter* λ *using the polynomial model and data sets from the course website as in Ex. 1 of sheet 3.*
Optimize λ *using leaving-one-out cross-validation (LOOCV). Does LOOCV find the parameter value giving the lowest testing error?*

3 points

N. Bertschinger
M. Kaschube
V. Ramesh

Exercise 3. *Implement Bayesian linear regression as explained in the lecture:*

- *Prior:* $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$
- *Likelihood:* $t \sim \mathcal{N}(\mathbf{w}^T \mathbf{\Phi}(x), \beta^{-1})$
- *Posterior (on training data* $D = \{(x_n, t_n)\}_{n=1}^N$*):*

$$p(\mathbf{w}|D) = \mathcal{N}(\mu_N, \Sigma_N)$$

where

$$\begin{aligned} \Sigma_N &= (\alpha \mathbf{I} + \beta \mathbf{\Phi}^T \mathbf{\Phi})^{-1} \\ \mu_N &= \beta \Sigma_N \mathbf{\Phi}^T \mathbf{t} \end{aligned}$$

with design matrix $(\mathbf{\Phi})_{ni} = (\mathbf{\Phi}(\mathbf{x_n}))_i$.

- *Evidence:*

$$\log p(D) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma_N| - \frac{\alpha}{2} \mu_N^T \mu_N - \frac{\beta}{2} ||\mathbf{t} - \mathbf{\Phi} \mathbf{w}||^2$$

1. *Illustrate how the marginal likelihood can be used for model selection on the polynomial regression example, again using the same data sets from the course website.*
2. *Compare your results to the ones obtained with LOOCV in Ex. 2 above.*

5 points