N. Bertschinger
M. Kaschube
V. Ramesh

Machine Learning I
2. Exercise Sheet

Exercise 1. *Consider a data set in which each data point* $(\mathbf{x}_n, t_n)$ *is associated with a weighting factor* $r_n > 0$*, so that the sum-of-squares error function becomes*

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_n r_n(t_n - \mathbf{w}^T\Phi(\mathbf{x_n}))^2$$

Find an expression for the solution $\mathbf{w}^*$ *that minimizes this error function. Give two alternative interpretations of the weighted sum-of-squares error function in terms of (i) data dependent noise variance and (ii) replicated data points.* **2 points**

Exercise 2. *Consider a linear model of the form*

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

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

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

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* $\mathbb{E}[\epsilon_i\epsilon_j] = \delta_{ij}\sigma^2$ *, 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.* **3 points**

Exercise 3. *Consider a standard Gaussian distribution in D dimensions:*

$$p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}||x||^2}$$

We wish to find the density with respect to radius in polar coordinates in which the direction variables have been integrated out. To do this, show that the integral of the probability density over a thin shell of radius r and thickness ϵ*, where* $\epsilon \ll 1$*, is given by*

$$p(r) = S_D r^{D-1}(2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}r^2}$$

Here, S_D *denotes the surface area of a unit sphere in D dimensions.*
Show that $p(r)$ *has a maximum at* $r^* = \sqrt{D-1}$*, while the density of* $\mathbf{x}$ *at this distance to the origin, i.e.* $||\mathbf{x}|| = r$ *is smaller than* $p(\mathbf{X} = \mathbf{0})$ *by a factor of* $e^{\frac{D}{2}}$*.* **3 points**

N. Bertschinger
M. Kaschube
V. Ramesh

Exercise 4. *Cross-validation can be considered as an estimator for the "true" expected loss* $\mathbb{E}[l]$ *of a model.*
Explain why it provides a better estimate of the expected loss, than the training set loss, i.e. $\min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} l(t_n, y(x_n; \mathbf{w}))$.
Illustrate that there is a bias-variance tradeoff in choosing the number of folds in cross-validation. **2 points**