Machine Learning 1 - Homework assignment 2

Available: Monday, November 5th, 2018
Deadline: Thursday 23.59, November 15th, 2018

General instructions
Unless stated otherwise, write down a derivation of your solutions. Solutions presented without a derivation that shows how the solution was obtained will not be awarded with points.

1 MAP solution with correlated responses

In practice assignment 2 you derived the maximum a posterior for linear regression assuming *independent responses*. In general, it is possible that a measurement device which is used to record data is influenced by the measurement process itself, ultimately leading to *correlated measurements*. We will therefore drop the assumption of independence and assume correlated measurements, quantified by a known, nonsingular covariance matrix $\mathbf{\Omega}$.

As before, we assume N training vectors $\{\mathbf{x}_n\}_{n=1}^N$, each of which is mapped to a different feature vector $\boldsymbol{\phi}_n = (\phi_0(\mathbf{x}_n), \phi_1(\mathbf{x}_n), ..., \phi_{M-1}(\mathbf{x}_n))^T$ using basis functions $\phi_j(\mathbf{x})$ with $j = 0, ..., M-1$ and a training set consisting of tuples $(\mathbf{x}_n, t_n)$. The model assumptions are altered to:

- The regression prediction is given by: $y(\mathbf{x}_n, \mathbf{w}) = \mathbf{w}^T \boldsymbol{\phi}_n$.
- The data samples are *not* longer i.i.d.
- The likelihood function *of the whole dataset* is a Gaussian: $p(\mathbf{t}|\mathbf{\Phi}, \mathbf{w}, \mathbf{\Omega}) = \mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w}, \mathbf{\Omega})$, where $\mathbf{\Omega}$ is the covariance matrix of the measurements with entries $\Omega_{ij} = \mathrm{Cov}(t_i, t_j)$.
- The prior over $\mathbf{w}$ is given by: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$, where $\mathbf{0}$ is a vector of 0's.

Derive the MAP solution

$$\mathbf{w}_{\mathrm{MAP}} = \left(\alpha\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Omega}^{-1}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{\Omega}^{-1}\mathbf{t}$$

for this problem by reducing it to the uncorrelated case. The decorrelation is done via a change to the eigenbasis of the covariance matrix.

a) Write down the likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ in vector/matrix form, i.e. in terms of $\mathbf{t}$, $\boldsymbol{\Phi}$, $\mathbf{w}$ and $\boldsymbol{\Omega}$. Note that the distribution can not be factored into independent multiplicands in this basis.

b) Write the likelihood in terms of a Gaussian distribution with a diagonal covariance matrix by changing the basis of the space in which the targets are expressed. Specifically, express the covariance matrix in its eigenbasis, i.e. write it as $\boldsymbol{\Omega} = \mathbf{A}^T\mathbf{D}\mathbf{A}$ with $\mathbf{D} := \mathrm{diag}(d_1, \ldots, d_N)$ being a diagonal matrix containing the eigenvalues of $\boldsymbol{\Omega}$ and $\mathbf{A}^T = \mathbf{A}^{-1}$ being an *orthogonal* change of basis. This is possible in general since covariance matrices are symmetric[1].
Hint: Substitute $\boldsymbol{\tau} := \mathbf{A}\mathbf{t}$ *and* $\boldsymbol{\Psi} := \mathbf{A}\boldsymbol{\Phi}$ *in the exponent and use the properties* $\det(\mathbf{U}\mathbf{V}) = \det(\mathbf{U})\det(\mathbf{V})$ *and* $\det(\mathbf{U}^{-1}) = \det(\mathbf{U})^{-1}$ *of the determinant in the normalization constant.*

c) Factorize the distribution, into a product of univariate Gaussians. Write ψ_i for the i-th row of the matrix $\boldsymbol{\Psi}$.
Hint: The determinant of a diagonal matrix can be expressed as the product of its elements.

d) Write down the explicit form of the prior $p(\mathbf{w})$, i.e. use the expression for a multivariate Gaussian distribution with the correct mean and covariance. Compute the logarithm of the prior $\ln p(\mathbf{w})$.

e) Write down an expression for the posterior $p(\mathbf{w}|\mathcal{D})$ over $\mathbf{w}$ by applying Bayes rule. You do not need to write out the explicit form of the Gaussian distributions, instead use the form $\mathcal{N}(a|b, c^2)$ with appropriate means b and variances c^2. Show that the evidence will require an integral, which you do not need to solve analytically! However, you need to replace it with a probability distribution like $p(a|b, c)$ with the correct corresponding variables and conditioning variables. (Note that $p(a|b, c)$ is just an example, there might be more or less than 2 conditioning variables.)

f) Compute the log-posterior for both the matrix form of the likelihood as derived in 1.2 and the factorized component form of the likelihood as derived in 1.3. Collect all terms which are independent of $\mathbf{w}$ into a constant C. Which parts of the previous expression do not depend on $\mathbf{w}$? Why is finding the MAP much simpler than finding the full posterior distribution?

[1]For more information see `https://en.wikipedia.org/wiki/Spectral_theorem`

g) Solve for $\mathbf{w}_{\text{MAP}}$ by first taking the derivative of the log-posterior with respect to $\mathbf{w}$, then setting it to 0, and finally solving for $\mathbf{w}$.

h) Express the solution for $\mathbf{w}_{\text{MAP}}$ in terms of the original quantities $\mathbf{t}$ and $\mathbf{\Phi}$ to end up with the final solution stated above.

2 ML estimate of angle measurements

Find the maximum likelihood estimate of the angle θ, given that you have two independent noisy measurements, c and s, where c is a measure of the cosine, and s is a measure of the sine, of the angle θ. Assume each measurement has a known Gaussian standard deviation σ (same in both cases).
Hint: Use that $\sin^2\theta + \cos^2\theta = 1\ \forall\theta \in [0, 2\pi)$ *and gather all terms which are independent of* θ *into a constant* C*.*

3 ML and MAP solution of a Poisson distribution fit

Suppose we have a set of N integer non-negative measurements, $\{x_i\}_{i=1}^N$ from a stochastic process.

a) Demonstrate analytically that the maximum likelihood fit of a Poisson distribution to this data is given by $\lambda_{\text{ML}} = \frac{1}{N}\sum_{i=1}^N x_i$.

b) You are then given additional prior information that $p(\lambda) \propto \exp(-\lambda/a)$ for $\lambda \geq 0$ and zero otherwise, for a known value $a > 0$. What is the maximum posterior solution for λ?

c) What is the effect of the prior on the value of the estimate, (compared to the maximum likelihood case) and what happens in the limits $a \to \infty$ and $a \to 0$?