Machine Learning 1 - Practice exercise 2

1 MAP solution for Linear Regression

In class we solved for the maximum likelihood estimator for linear regression with basis functions. In this exercise you will solve for the *maximum a posterior* (MAP) solution. For this problem 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 $j = 0, ..., M-1$ where we define a bias $\phi_0(\mathbf{x}) = 1$. In the training set, the data come in input-output pairs: $(\mathbf{x}_n, t_n)$. We have the following model assumptions:

- The regression prediction is given by: $y(\mathbf{x}_n, \mathbf{w}) = \mathbf{w}^T \boldsymbol{\phi}_n$.
- The data samples are i.i.d. (independently and identically distributed).
- The likelihood function is a Gaussian: $p(t_n|\boldsymbol{\phi}_n, \mathbf{w}, \beta) = \mathcal{N}\left(t_n|\mathbf{w}^T\boldsymbol{\phi}_n, \beta^{-1}\mathbf{I}\right)$, where $\mathbf{I}$ is the identity matrix.
- The prior over $\mathbf{w}$ is given by: $p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right)$, where $\mathbf{0}$ is a vector of 0's.

The MAP solution for the weights $\mathbf{w}$ turns out to be given by $\mathbf{w}_{\text{MAP}} = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbf{I}\right)^{-1}\boldsymbol{\Phi}^T\mathbf{t}$, with

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \dots & \phi_{M-1}(\mathbf{x}_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \dots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

Derive the MAP solution by answering the following questions:

a) Write down the likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ using: a) a product over N, and b) in vector/matrix form.
Hint: You can answer both a) and b) in one set of equations by starting with a), then simplifying to get b). For b) make sure to define any matrices and vectors.

b) 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 $\ln p(\mathbf{w})$.

c) 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.)

d) Compute the log-posterior for both expressions for the likelihood from question 1.1.a) and 1.1.b). 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?

e) 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 $\mathbf{w}$. Do this for both forms of log-posterior that you wrote down in question 1.4.

f) Our prior for $\mathbf{w}$ assumes the same distribution for each entry in $\mathbf{w}$, including w_0 which is multiplied by the basis function $\phi_0(\mathbf{x}) = 1$ in the regression prediction function $y(\mathbf{x}, \mathbf{w})$. What is the role of w_0 and $\phi_0(\mathbf{x})$? Why should we avoid placing the same penalty/prior for this basis? Rewrite $p(\mathbf{w})$ so that w_0 has its own prior/penalty.

Solutions

a)

$$
\begin{aligned}
p(\mathcal{D}|\boldsymbol{\theta}) &= \prod_{n=1}^{N} \mathcal{N}\left(t_n|\mathbf{w}^T\boldsymbol{\phi}_n, 1/\beta\right) \\
&= \prod_{n=1}^{N} \frac{\beta^{1/2}}{(2\pi)^{1/2}} \exp\left(-\frac{\beta}{2}\left(t_n - \mathbf{w}^T\boldsymbol{\phi}_n\right)^2\right) \\
&= \frac{\beta^{N/2}}{(2\pi)^{N/2}} \prod_{n=1}^{N} \exp\left(-\frac{\beta}{2}\left(t_n - \mathbf{w}^T\boldsymbol{\phi}_n\right)^2\right) \\
&= \frac{\beta^{N/2}}{(2\pi)^{N/2}} \exp\left(-\frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^T\boldsymbol{\phi}_n\right)^2\right) \\
&= \frac{\beta^{N/2}}{(2\pi)^{N/2}} \exp\left(-\frac{\beta}{2}\left(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\right)^T\left(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\right)\right) \\
&= \mathcal{N}\left(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, \frac{1}{\beta}\mathbf{I}\right)
\end{aligned}
$$

b)

$$
\begin{aligned}
p(\mathbf{w}) &= \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \frac{1}{\alpha}\mathbf{I}\right) \\
&= \frac{\alpha^{D/2}}{(2\pi)^{D/2}} \exp\left(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right)
\end{aligned}
$$

$$
\begin{aligned}
\ln p(\mathbf{w}) &= \frac{D}{2}\ln\alpha - \frac{D}{2}\ln(2\pi) - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \\
&= -\frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + C
\end{aligned}
$$

c)

$$
\begin{aligned}
p(\mathbf{w}|\mathcal{D}) &= \frac{\mathcal{N}\left(\mathbf{w}|\mathbf{0}, \frac{1}{\alpha}\mathbf{I}\right)\prod_{n=1}^{N}\mathcal{N}\left(t_n|\mathbf{w}^T\boldsymbol{\phi}_n, 1/\beta\right)}{\int \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \frac{1}{\alpha}\mathbf{I}\right)\prod_{n=1}^{N}\mathcal{N}\left(t_n|\mathbf{w}^T\boldsymbol{\phi}_n, 1/\beta\right) d\mathbf{w}} \\
&= \frac{\mathcal{N}\left(\mathbf{w}|\mathbf{0}, \frac{1}{\alpha}\mathbf{I}\right)\mathcal{N}\left(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, \frac{1}{\beta}\mathbf{I}\right)}{\int \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \frac{1}{\alpha}\mathbf{I}\right)\mathcal{N}\left(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, \frac{1}{\beta}\mathbf{I}\right) d\mathbf{w}} \\
&= \frac{\mathcal{N}\left(\mathbf{w}|\mathbf{0}, \frac{1}{\alpha}\mathbf{I}\right)\mathcal{N}\left(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, \frac{1}{\beta}\mathbf{I}\right)}{p(\mathbf{t}|\boldsymbol{\Phi}, \alpha, \beta)}
\end{aligned}
$$

d)

$$\begin{aligned}\ln p(\mathbf{w}|\mathcal{D}) &= -\frac{\alpha}{2}\mathbf{w}^T\mathbf{w} - \frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^T\boldsymbol{\phi}_n\right)^2 + C \\ &= -\frac{\alpha}{2}\mathbf{w}^T\mathbf{w} - \frac{\beta}{2}\left(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\right)^T\left(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\right) + C\end{aligned}$$

e) In index notation:

$$\ln p(\mathbf{w}|\mathcal{D}) = -\frac{\alpha}{2}\mathbf{w}^T\mathbf{w} - \frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^T\boldsymbol{\phi}_n\right)^2 + C$$

$$\frac{\partial \ln p(\mathbf{w}|\mathcal{D})}{\partial \mathbf{w}} = -\alpha\mathbf{w}^T - \beta\sum_{n=1}^{N}\left(t_n - \mathbf{w}^T\boldsymbol{\phi}_n\right)\left(-\boldsymbol{\phi}_n^T\right) = 0$$

Taking the transpose of both sides:

$$\begin{aligned}\alpha\mathbf{w} &= \beta\sum_{n=1}^{N}\boldsymbol{\phi}_n\left(t_n - \boldsymbol{\phi}_n^T\mathbf{w}\right) \\ \alpha\mathbf{w} &= \beta\sum_{n=1}^{N}t_n\boldsymbol{\phi}_n - \boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\mathbf{w} \\ \left(\alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\right)\mathbf{w} &= \beta\sum_{n=1}^{N}t_n\boldsymbol{\phi}_n \\ \mathbf{w}_{\text{MAP}} &= \left(\alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\right)^{-1}\beta\sum_{n=1}^{N}t_n\boldsymbol{\phi}_n\end{aligned}$$

In matrix form:

$$\begin{aligned}\ln p(\mathbf{w}|\mathcal{D}) &= -\frac{\alpha}{2}\mathbf{w}^T\mathbf{w} - \frac{\beta}{2}\left(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\right)^T\left(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\right) + C \\ &= -\frac{\alpha}{2}\mathbf{w}^T\mathbf{w} - \frac{\beta}{2}\mathbf{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w} + \beta\mathbf{w}^T\boldsymbol{\Phi}^T\mathbf{t} + D\end{aligned}$$

$$\frac{\partial \ln p(\mathbf{w}|\mathcal{D})}{\partial \mathbf{w}} = -\alpha\mathbf{w}^T - \beta\mathbf{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \beta\mathbf{t}^T\boldsymbol{\Phi} = 0$$

Taking the transpose of both sides:

$$\begin{aligned}
\left(\alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}\right)\mathbf{w} &= \beta\mathbf{\Phi}^T\mathbf{t} \\
\mathbf{w}_{\text{MAP}} &= \left(\alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\beta\mathbf{\Phi}^T\mathbf{t} \\
&= \left(\beta\left(\frac{\alpha}{\beta}\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Phi}\right)\right)^{-1}\beta\mathbf{\Phi}^T\mathbf{t} \\
&= \frac{1}{\beta}\left(\lambda\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\beta\mathbf{\Phi}^T\mathbf{t} \\
&= \left(\lambda\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{t}
\end{aligned}$$

f) The constant basis function acts as a bias or offset for the regression problem. If we use the same prior for this weight as for the others, we are assuming that the offset from the y-axis should somehow be penalized. This does not make too much sense a priori, so instead we use a different precision for this basis function, i.e. $\alpha_0 << \alpha$, while using α for all the others.

2 Probability distributions, likelihoods, and estimators

For these questions you will be working with different probability density functions (PDFs) listed in the table below. The purpose of these questions is to practice working with a variety of PDFs and to make computing likelihoods, maximum likelihood (ML) estimates, etc. more natural. Note below the *indicator* notation $[x = 0]$ (and $[x = 1]$). The square brackets evaluate to 1 if the argument is true, and 0 otherwise. E.g. if x is 1, the $[x = 0] = 0$ and $[x = 1] = 1$ (here $[x = 0]$ is lazy notation; in Python you would write $x == 0$, for example). We will use this notation of an indicator function a lot, both below and when we learn about classification.

Distribution	$p(x\|\theta)$	Range of x	Range of θ
Bernouilli	$\theta^{[x=1]}(1-\theta)^{[x=0]}$	$x \in \{0, 1\}$	$0 \leq \theta \leq 1$
Beta	$\frac{\Gamma(\theta_1+\theta_0)}{\Gamma(\theta_1)\Gamma(\theta_0)}x^{\theta_1-1}(1-x)^{\theta_0-1}$	$0 \leq x \leq 1$	$\theta_1 > 0, \theta_0 > 0$
Poisson	$\frac{\theta^x}{x!}e^{-\theta}$	$x \in \{0, 1, 2, \ldots\}$	$\theta > 0$
Gamma	$\frac{\theta_1^{\theta_0}}{\Gamma(\theta_0)}x^{\theta_0-1}e^{-\theta_1 x}$	$x > 0$	$\theta_1 > 0$, $\theta_0 > 0$
Gaussian	$\frac{1}{\sqrt{2\pi\theta_1}}e^{-\frac{1}{2}\left(\frac{x-\theta_0}{\theta_1}\right)^2}$	$-\infty < x < \infty$	$-\infty < \theta_0 < \infty$, $\theta_1 > 0$
Log-Normal	$\frac{1}{x\sqrt{2\pi\theta_1}}e^{-\frac{1}{2}\left(\frac{\ln x-\theta_0}{\theta_1}\right)^2}$	$x > 0$	$-\infty < \theta_0 < \infty$, $\theta_1 > 0$

Question 2.1

For each of the distributions in the table describe in your own words what kind of events and variables the distribution at hand can model and provide one example.

Question 2.2

You live in Den Helder and find that it rains quite a lot. Your goal is to estimate the probability that it will rain on *any* given day of the year. For one year, for each month, you count the number of days with rain. You get the following counts (from January to December):

$$21, 18, 17, 15, 13, 12, 15, 15, 18, 2, 21, 22$$

(for a grand total of 207 days with rain)[1].

Let r_t be a binary random variable denoting the observation for day t in that year; $r_t = 1$ means it rained on day t, and $r_t = 0$ means it did not rain. We want to estimate the probability, ρ, of rain on any day of the year. To answer these questions, the number of days of rain per month is not important, only the total for the year is relevant. With this information, answer the following questions:

a) What is the likelihood for a single observation r_t? And what is the likelihood for the entire set of observations $\{r_t\}_{t=1}^{N}$? Use n_1 to indicate the total number of days of rain, and n_0 to indicate the total number of days without rain, and N for the total number of days.

b) Write the log-likelihood for the entire set of observations.

c) Solve for the maximum likelihood (ML) estimate of ρ. Do it in general (with symbols for counts n_0, n_1 for days without and with rain) and for this specific case (plug-in the numbers).

d) Assume a Beta prior for ρ with parameters a and b. Solve for the MAP estimate for ρ.

e) Write the form of the posterior distribution for ρ? You do not need to solve it analytically.

f) (Bonus) Solve for the posterior distribution analytically. Hint: it is a Beta distribution.

[1]Source: http://www.amsterdam.climatemps.com/.

Question 2.3

You work in the staffing department of a maternity hospital and part of your job is to determine the staffing requirements during the night shift at your hospital. This might mean the number of doctors and nurses at the hospital and the number of doctors on call (if there are more than the average number of deliveries). Your goal is to determine the distribution over the number of deliveries during the night shift $d_t \in \{0, 1, 2, \ldots\}$ (d for delivery count, t for time, the index of the night). With this you can compute the mean, the probability of more than 5 deliveries, etc. You collect data for two weeks, i.e. $d_1, \ldots, d_{14} = 4, 7, 3, 0, 2, 2, 1, 5, 4, 4, 3, 3, 2, 3$. You assume the observations are explained by a Poisson distribution with parameter λ over the discrete delivery counts. With this information, answer the following questions:

a) What is the likelihood for a single observation (in general)? What is the likelihood for the entire set of observations (in general)? For both questions use T to denote the total number of observations, and n for the total number of deliveries.

b) Compute the log-likelihood for the entire set of observations. Do not plug in the actual numbers, keep using the general symbols d_t, n and T.

c) Solve for the ML estimate of λ. Do it both for the general case and for the specific case mentioned in the introduction (plug-in the numbers).

d) Assume a Gamma prior for λ with parameters a and b. Compute the MAP estimate of λ for the general case.

e) Write down the form of the posterior distribution for λ. You do not need to solve it analytically.

f) (Bonus) Solve for the posterior distribution analytically. Hint: it is a Gamma distribution.

Solutions

Question 2.1

Bernouilli: binary probability distributions,a single bit of information whose value is success/yes/true/one with probability p and failure/no/false/zero with probability q

Beta: continuous probability distributions, conjugate prior probability distribution for the Bernoulli, binomial, negative binomial, applied to model the behavior of random variables limited to intervals of finite length

Poisson: discrete probability distribution, probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event

Gamma: continuous probability distribution,a conjugate prior distribution for various types of inverse scale (aka rate) parameters, such as the λ of an exponential distribution or a Poisson distribution

Gaussian: continuous probability distribution, used in the natural and social sciences to represent real-valued random variables whose distributions are not known

Log-Normal: continuous probability distribution, a random variable whose logarithm is normally distributed

Question 2.2 a)

$$p(r_t|\rho) = \rho^{r_t}(1-\rho)^{1-r_t}$$

$$\begin{aligned} p(\mathbf{r}|\rho) &= \prod_{t=1}^{T} \rho^{r_t}(1-\rho)^{1-r_t} \\ &= \rho^{\sum_{t=1}^{T} r_t}(1-\rho)^{\sum_{t=1}^{T} 1-r_t} \\ &= \rho^{n_1}(1-\rho)^{n_0} \end{aligned}$$

b)

$$\ln p(\mathbf{r}|\rho) = n_1 \ln \rho + n_0 \ln(1-\rho)$$

c)

$$f = \ln p(\mathbf{r}|\rho) = n_1 \ln \rho + n_0 \ln(1-\rho)$$

$$\begin{aligned}
\partial f/\partial \rho &= \frac{n_1}{\rho} + \frac{n_0}{1-\rho}(-1) = 0 \\
\frac{n_1}{\rho} &= \frac{n_0}{1-\rho} \\
n_1 - n_1\rho &= n_0\rho \\
\rho &= \frac{n_1}{N} \\
&= 207/365
\end{aligned}$$

d)

$$\begin{aligned}
f = \ln p(\rho|\mathbf{r}) &\propto \ln p(\mathbf{r}|\rho) + \ln p(\rho) \\
&= n_1 \ln \rho + n_0 \ln(1-\rho) + (a-1)\ln\rho + (b-1)\ln(1-\rho)
\end{aligned}$$

$$\begin{aligned}
\partial f/\partial \rho &= \frac{n_1}{\rho} - \frac{n_0}{1-\rho} + \frac{a-1}{\rho} - \frac{b-1}{1-\rho} = 0 \\
\rho &= \frac{n_1 + a - 1}{N + a + b - 2}
\end{aligned}$$

e)

$$\begin{aligned}
p(\rho|\mathbf{r}) &= p(\mathbf{r}|\rho)p(\rho) \\
&= \frac{\rho^{n_1+a-1}(1-\rho)^{n_0+b-1}}{\int \rho^{n_1+a-1}(1-\rho)^{n_0+b-1} d\rho} \\
&= \frac{\Gamma(N+a+b)}{\Gamma(n_1+a)\Gamma(n_0+b)} \rho^{n_1+a-1}(1-\rho)^{n_0+b-1} \\
&= \mathcal{B}(\rho|a+n_1, b+n_0) \quad \text{Beta distribution}
\end{aligned}$$

f) See previous answer.

Question 2.3 a)

$$p(d_t|\lambda) = \frac{\lambda^{d_t}}{d_t!}\exp(-\lambda)$$

$$\begin{aligned} p(\mathbf{d}|\lambda) &= \prod_{t=1}^{T} \frac{\lambda^{d_t}}{d_t!}\exp(-\lambda) \\ &= \frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t!}\exp(-T\lambda) \\ &= \frac{\lambda^{n}}{\prod_{t=1}^{T} d_t!}\exp(-T\lambda) \end{aligned}$$

b)

$$\ln p(\mathbf{d}|\lambda) = n\ln\lambda - T\lambda - \sum_{t=1}^{T}\ln(d_t!)$$

c)

$$f = \ln p(\mathbf{d}|\lambda) = n\ln\lambda - T\lambda - \sum_{t=1}^{T}\ln(d_t!)$$

$$\begin{aligned} \partial f/\partial\lambda &= n/\lambda - T = 0 \\ \lambda &= n/T = 43/14 \end{aligned}$$

d)

$$f = \ln p(\mathbf{d}|\lambda) + \ln p(\lambda) = n\ln\lambda - T\lambda - \sum_{t=1}^{T} d_t! + (a-1)\ln\lambda - b\lambda + C$$

$$\begin{aligned} \partial f/\partial\lambda &= n/\lambda - T + (a-1)/\lambda - b = 0 \\ \lambda &= \frac{n+a-1}{T+b} \end{aligned}$$

e)

$$
\begin{aligned}
p(\lambda|\mathbf{d}) &= \frac{p(\mathbf{d}|\lambda)p(\lambda)}{\int p(\mathbf{d}|\lambda)p(\lambda)d\lambda} \\
&= \frac{\frac{\lambda^n}{\prod_{t=1}^{T} d_t!}\exp(-T\lambda)\frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda)}{\int \frac{\lambda^n}{\prod_{t=1}^{T} d_t!}\exp(-T\lambda)\frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda)d\lambda} \\
&= \frac{\lambda^{n+a-1}\exp(-(T+b)\lambda)}{\int \lambda^{n+a-1}\exp(-(T+b)\lambda)d\lambda} \\
&= \frac{(T+b)^{n+a}}{\Gamma(n+a)}\lambda^{n+a-1}\exp(-(T+b)\lambda) \\
&= \mathcal{G}\left(\lambda|a+n, b+T\right) \quad \text{Gamma distribution}
\end{aligned}
$$

f) See previous answer.