— Solution notes —

Second practicals in Machine learning 1 – 2024 – Paper 2

1 MAP solution for linear regression (September)

In this exercise, you will solve and practice with the maximum a posterior (MAP) estimator for linear regression with basis function. 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: $(\mathbf{x_n}, t_n)$. Moreover, 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: $p(\mathbf{t}|\mathbf{\Phi}, \mathbf{w}, \beta) = \mathcal{N}(\mathbf{t}|\mathbf{\Phi w}, \beta^{-1}\mathbf{I})$ where I is the identity matrix. 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.

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

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

Derive the MAP solution by answering the following questions:

(a) Write down the likelihood $p(\mathcal{D}|\mathbf{w})$ 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.

Answer:

$$p(\mathcal{D}|\mathbf{w}) \overset{\{i.i.d.\}}{=} \prod_{n=1}^{N} \mathcal{N}(t_n|w^T\boldsymbol{\phi}_n, 1/\beta) = \prod_{n=1}^{N} \frac{\beta^{1/2}}{(2\pi)^{1/2}} exp\left(-\frac{\beta}{2}(t_n - \mathbf{w}^T\boldsymbol{\phi_n})^2\right)$$

$$= \frac{\beta^{N/2}}{(2\pi)^{N/2}} \prod_{n=1}^{N} exp\left(-\frac{\beta}{2}(t_n - \mathbf{w}^T\boldsymbol{\phi_n})^2\right) = \frac{\beta^{N/2}}{(2\pi)^{N/2}} exp\left(-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\boldsymbol{\phi_n})^2\right)$$

Finally, we can vectorize the final expression as such:

$$\Leftrightarrow p(\mathcal{D}|\mathbf{w}) = \frac{\beta^{N/2}}{(2\pi)^{N/2}} exp\left(-\frac{\beta}{2}(\mathbf{t} - \mathbf{\Phi w})^T(\mathbf{t} - \mathbf{\Phi w})\right) = \mathcal{N}\left(\mathbf{t}|\mathbf{\Phi w}, \frac{1}{\beta}I\right)$$

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

Answer:

$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \frac{1}{\alpha}I\right)$$

$$= \frac{\alpha^{M/2}}{(2\pi)^{M/2}} exp\left(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right)$$

$$\Leftrightarrow ln(p(\mathbf{w})) = \frac{M}{2}ln(\alpha) - \frac{M}{2}ln(2\pi) - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

$$= C - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

(*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 $N(a|b, c^2)$ 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.

Answer:

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

$$= \frac{\mathcal{N}\left(\mathbf{w}|\mathbf{0}, \frac{1}{\alpha}I\right)\mathcal{N}\left(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, 1/\beta I\right)}{\int \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \frac{1}{\alpha}I\right)\mathcal{N}\left(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, 1/\beta I\right) dw}$$

$$= \frac{\mathcal{N}\left(\mathbf{w}|\mathbf{0}, \frac{1}{\alpha}I\right)\mathcal{N}\left(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, 1/\beta I\right)}{p(\mathbf{t}|\boldsymbol{\Phi}, \alpha, \beta)}$$

(*d*) Compute the log-posterior for both expressions for the likelihood from question 1.a) and 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?

Answer:

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

$$= -\frac{\alpha}{2}\mathbf{w}^T\mathbf{w} - \frac{\beta}{2}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})^T(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}) + C$$

(*e*) Solve for $\mathbf{w_{MAP}}$ 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.d.

Answer: Firstly regarding the index notation:

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

By using the chain rule:

$$\frac{dln(p(\mathbf{w}|\mathcal{D}))}{d\mathbf{w}} = -\alpha\mathbf{w}^T - \beta\sum_{n=1}^{N}(t_n - \mathbf{w^T}\boldsymbol{\phi_n})(-\boldsymbol{\phi_n}^T) = 0$$

$$\Leftrightarrow \alpha\mathbf{w}^T = \beta\sum_{n=1}^{N}(t_n - \mathbf{w^T}\boldsymbol{\phi_n})(\boldsymbol{\phi_n}^T)$$

By taking the transpose in both sides:

$$\alpha\mathbf{w} = \beta\sum_{n=1}^{N}\boldsymbol{\phi_n}(t_n - \boldsymbol{\phi_n}^T\mathbf{w})$$

Then, we can expand the summation into:

$$\Leftrightarrow \alpha\mathbf{w} = \beta\sum_{n=1}^{N}\boldsymbol{\phi_n}t_n - \beta\sum_{n=1}^{N}\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\mathbf{w}$$

Collect all the terms that depends on $\mathbf{w}$:

$$\Leftrightarrow \alpha\mathbf{w} + \beta\sum_{n=1}^{N}\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\mathbf{w} = \beta\sum_{n=1}^{N}t_n\boldsymbol{\phi_n}$$

$$\Leftrightarrow \left(\alpha I + \beta\sum_{n=1}^{N}\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\right)\mathbf{w} = \beta\sum_{n=1}^{N}t_n\phi_n$$

Finally, by multiplying the inverse $\left(\alpha I + \beta\sum_{n=1}^{N}\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\right)^{-1}$ in both sides of the equation we derive to the following solution:

$$\Leftrightarrow \mathbf{w_{MAP}} = \left(\alpha I + \beta\sum_{n=1}^{N}\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\right)^{-1}\beta\sum_{n=1}^{N}t_n\boldsymbol{\phi_n}$$

In matrix form:

$$lnp(\mathbf{w}|\mathcal{D}) = -\frac{\alpha}{2}\mathbf{w}^T\mathbf{w} - \frac{\beta}{2}\mathbf{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w} + \beta\mathbf{w}^T\boldsymbol{\Phi}^T\mathbf{t} + D$$

$$\frac{dlnp(\mathbf{w}|\mathcal{D})}{d\mathbf{w}} = -\alpha\mathbf{w}^T - \beta\mathbf{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \beta\mathbf{t}^T\boldsymbol{\Phi}$$

...

$$\Leftrightarrow \mathbf{w}_{MAP} = (\alpha I + \beta \mathbf{\Phi^T \Phi})^{-1} \beta \mathbf{\Phi^T t}$$

$$\Leftrightarrow \mathbf{w}_{MAP} = (\frac{\alpha}{\beta} I + \mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{t}$$

(*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 $\mathbf{w_0}$ 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.

Answer: 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 penalised. 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.

Second practicals in Machine learning 1 – 2024 – Paper 2

2 Probability distributions, likelihoods, and estimators exercise 3 (September)

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.

Answer: The likelihood for a single observation is given by:

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

While the likelihood for the entire set:

$$p(\mathbf{r}|\rho) = \prod_{t=1}^{T} \rho^{r_t}(1-\rho)^{(1-r_t)}$$

$$\Leftrightarrow p(\mathbf{r}|\rho) = \rho^{\sum t=1^{T} r_t}(1-\rho)^{\sum t=1^{T}(1-r_t)}$$

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

Answer:

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

(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).

Answer:

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

.

By taking the derivative over ρ and setting it into zero:

$$\frac{\partial f}{\partial \rho} = \frac{n_1}{\rho} + \frac{n_0}{1-\rho}(-1) = 0$$

.

$$\Leftrightarrow \frac{n_1}{\rho} = \frac{n_0}{1-\rho}$$

Hence, the ML estimate is found to be:

$$\Leftrightarrow \rho_{ML} = \frac{207}{365}$$

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

Answer:

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

Since the evidence does not depend on ρ we can write the following:

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

$$\frac{\partial f}{\partial \rho} = \frac{n_1}{\rho} - \frac{n_0}{1-\rho} + \frac{a-1}{\rho} - \frac{b-1}{1-\rho} = 0$$

Hence, the MAP estimate is found to be:

$$\rho_{MAP} = \frac{n_1 + a - 1}{N + a + b - 2}$$

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

Answer:

$$p(\rho|\mathbf{r}) = \frac{p(\mathbf{r}, \rho)}{p(\mathbf{r})} = \frac{p(\mathbf{r}|\rho)p(\rho)}{p(\mathbf{r})}$$

Then, we plug in the likelihood, prior, and by marginalizing out ρ to calculate the evidence, we can show that it is indeed a Beta distribution:

$$\mathcal{B}(\rho|a+n_1, b+n_0)$$

Beta distribution.

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

Second assignment in Machine learning 1 – 2024 – Paper 2

3 Maximum likelihood estimate of angle measurements (September)

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). To make the solution unique assume $\theta \in [-\pi/2, \pi/2]$.

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.

(*a*) Write down the likelihood $p(s, c|\theta)$. [1 point]

Answer: The independence of the measurements of s and c implies:

$$p(s, c|\theta) = p(s|\theta)p(c|\theta)$$

Correctly plugging in the functions and factorising terms:

$$\begin{aligned}
&\propto \exp\left(-\frac{1}{2\sigma^2}(s - \sin\theta)^2\right)\exp\left(-\frac{1}{2\sigma^2}(c - \cos\theta)^2\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left((s - \sin\theta)^2 + (c - \cos\theta)^2\right)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\underbrace{s^2 + c^2 + 1}_{:=C} - 2s\sin\theta - 2c\cos\theta\right)\right)
\end{aligned}$$

The likelihood can be expressed as following:

$$\propto \exp\left(\frac{1}{\sigma^2}\left(s\sin\theta + c\cos\theta\right)\right)$$

(*b*) Write down the log-likehood. [1 point]

Answer: The log-likelihood reads $\ln p(s, c|\theta) = \frac{1}{\sigma^2}(s\sin\theta + c\cos\theta)$.

(*c*) Obtain the maximum likelihood estimation. [1 point]

Answer:

To obtain the maximum likelihood solution, take the derivative w.r.t. θ and set it to zero:

$$\frac{\partial}{\partial\theta}\ln p(s, c|\theta) = \frac{1}{\sigma^2}(s\cos\theta - c\sin\theta) = 0$$

Correctly rearranging variable using the appropriate trigonometric identities:

$$s\cos\theta - c\sin\theta = 0$$

$$
\begin{aligned}
s\cos\theta &= c\sin\theta \\
\frac{s}{c}\cos\theta &= \sin\theta \\
\frac{s}{c} &= \frac{\sin\theta}{\cos\theta} \\
\frac{s}{c} &= \tan\theta
\end{aligned}
$$

Correct expression for maximum likelihood estimation:

$$\theta_{\mathrm{ML}} = \arctan\left(\frac{s}{c}\right)$$
