Machine Learning 1 HW 1 Solutions

September 11, 2015

1 Probability Theory

For these questions you will practice manipulating probabilities and probability density functions using the sum and product rules.

Question 1.1
Being a student in the Netherlands, you spend all your time in the cities of Amsterdam and Rotterdam. Based on your experience, the weather in Amsterdam is much nicer: the probability that it rains when you are in Amsterdam is 0.5, while the probability that it rains when in Rotterdam is 0.75. Amsterdam is where you spend most of your time: at any given moment, the probability that you are in Amsterdam is 0.8 and the probability that you are in Rotterdam is 0.2.

Based on the above:

1. Define the random variables and the values they can take on, both with symbols and numerically.

2. What is the probability that it does not rain when you are in Rotterdam?

   Let $W \in \{s, r\}$ be the weather (sunny or rainy) and $L \in \{r, a\}$ be the location (Rotterdam or Amsterdam). Then $P(W = s|L = r) = 1 - 0.75 = 0.25$.

3. What is the probability that it rains where you are?

$$\begin{aligned} P(W = r) &= P(L = r)P(W = r|L = r) + P(L = a)P(W = r|L = a) \\ &= 0.2 \times 0.75 + 0.8 \times 0.5 \\ &= 0.55 \end{aligned}$$

4. You wake up on the sidewalk, after a night out which you can't remember anything about but which clearly was not such a great idea. You can't recognize your surroundings, but you must be either in Amsterdam or Rotterdam. It is raining. What is the probability that you are in Amsterdam?

$$\begin{aligned} P(L = a|W = r) &= \frac{P(L = a)P(W = r|L = a)}{P(W = r)} \\ &= \frac{0.8 \times 0.5}{0.55} \\ &= 0.727 \end{aligned}$$

Question 1.2

In a particular city with a population of 500000, it estimated that 500 people have cancer. There is a blood test that 99 times out of 100 correctly diagnoses cancer in patients. It also, unfortunately, misdiagnoses 5% of people that do not have cancer. With this information, answer the following questions:

1. What is p(cancer) and p(not cancer)?

   Let $D \in \{c, n\}$ (disease is cancer/not cancer), $B \in \{p, n\}$ (blood test is positive/negative). Then

$$\begin{aligned} P(D = c) &= 500/500000 \\ &= 1/1000 \\ P(D = n) &= 1 - P(D = c) \\ &= 999/1000 \end{aligned}$$

2. If a patient takes the blood test and it returns positive, what is the probability the patient has cancer?

$$\begin{aligned} P(D=c|B=p) &= \frac{P(D=c)P(B=p|D=c)}{P(B=p)} \\ &= \frac{\frac{1}{1000}\frac{99}{100}}{\frac{1}{1000}\frac{99}{100}+\frac{999}{1000}\frac{5}{100}} \\ &= \frac{99}{99+999*5} \\ &\approx 1/51 \end{aligned}$$

3. What are some of the assumptions we are implicitly making when answering this question?

   Most critically, that patients taking the test most likely have symptoms, so the population prior for cancer does not apply. Instead, something like $P(D|\text{has symptoms})$ would be more realistic.

Question 1.3

For this question you will write the expression for the posterior parameter distribution for a simple data problem. Assume we observe N univariate data points $\{x_1, x_2, \ldots, x_N\}$. Further, we assume that they are generated by a Gaussian distribution with known variance σ^2, but unknown mean μ. Assume a prior Gaussian distribution over the unknown mean, i.e. $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$. When answering these questions, use $\mathcal{N}(a|b, c^2)$ to indicate a Gaussian (normal) distribution over a with mean b and variance c^2.

1. Write down the general expression for a posterior distribution, using θ for the parameter, $\mathcal{D}$ for the data. Indicate the *prior*, *likelihood*, *evidence*, and *posterior*.

$$\begin{aligned} \underbrace{p(\theta|\mathcal{D})}_{\text{posterior}} &= \frac{p(\theta)p(\mathcal{D}|\theta)}{\int p(\theta)p(\mathcal{D}|\theta)d\theta} \\ &= \frac{\overbrace{p(\theta)}^{\text{prior}}\,\overbrace{p(\mathcal{D}|\theta)}^{\text{likelihood}}}{\underbrace{p(\mathcal{D})}_{\text{evidence}}} \end{aligned}$$

2. Write the posterior for this particular example. You do not need an analytic solution.

$$p(\theta|\mathcal{D}) = \frac{\mathcal{N}(\mu|\mu_0, \sigma_0^2)\prod_{n=1}^{N}\mathcal{N}(x_n|\mu, \sigma^2)}{\int \mathcal{N}(\mu|\mu_0, \sigma_0^2)\prod_{n=1}^{N}\mathcal{N}(x_n|\mu, \sigma^2)d\mu}$$

2 Basic Linear Algebra and Derivatives

If you have problems with this go see one of the TAs for help!

Question 2.1

Let $\mathbf{A} = \begin{bmatrix} 3 & 5 \\ 2 & 3 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 9 \\ 5 \end{bmatrix}$

1. Compute $\mathbf{Ab}$
2. Compute $\mathbf{b}^T\mathbf{A}$
3. What is the vector $\mathbf{c}$ for which $\mathbf{Ac} = \mathbf{b}$
4. What is $\mathbf{A}^{-1}$?
5. Verify that $\mathbf{A}^{-1}\mathbf{b} = \mathbf{c}$. Show that this must be the case.

Question 2.2

Find the gradient of the following functions

1. $x^2 + 2x + 3$
2. $(2x^3 + 1)^2$

Find the partial derivative of the following functions with respect to x,y, z

1. $f(x, y, z) = (x + 2y)^2 \sin(xy)$
2. $f(x, y, z) = 2\log(x + y^2 - z)$
3. $f(x, y, z) = \exp(x\cos(y + z))$

Question 2.3

The following questions are good practice in manipulating vectors and matrices and they are very important for solving for posterior distributions.

Given the following expression:

$$(\mathbf{x}-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})+(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\,\mathbf{S}^{-1}\,(\boldsymbol{\mu}-\boldsymbol{\mu}_0)$$

where $\mathbf{x}$, $\boldsymbol{\mu}$, $\boldsymbol{\mu}_0$ are vectors and $\boldsymbol{\Sigma}^{-1}$ and $\mathbf{S}^{-1}$ are symmetric, invertible matrices.

Answer the following questions:

1. Expand the expression and gather terms.

$$\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}^T\mathbf{S}^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}_0^T\mathbf{S}^{-1}\boldsymbol{\mu}_0-2\boldsymbol{\mu}^T\mathbf{S}^{-1}\boldsymbol{\mu}_0$$

2. Collect all the terms that depend on $\boldsymbol{\mu}$ and those that do not.

$$\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1}\right)\boldsymbol{\mu}-2\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{S}^{-1}\boldsymbol{\mu}_0\right)+\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}_0^T\mathbf{S}^{-1}\boldsymbol{\mu}_0$$

3. Take the derivative with respect to $\boldsymbol{\mu}$, set to 0, and solve for $\boldsymbol{\mu}$.

$$f(\boldsymbol{\mu})=\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1}\right)\boldsymbol{\mu}-2\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{S}^{-1}\boldsymbol{\mu}_0\right)+\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}_0^T\mathbf{S}^{-1}\boldsymbol{\mu}_0$$

$$\begin{aligned}\frac{\partial f(\boldsymbol{\mu})}{\partial\boldsymbol{\mu}}&=2\left(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1}\right)\boldsymbol{\mu}-2\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{S}^{-1}\boldsymbol{\mu}_0\right)\\\left(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1}\right)\boldsymbol{\mu}&=\left(\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{S}^{-1}\boldsymbol{\mu}_0\right)\\\boldsymbol{\mu}&=\left(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{S}^{-1}\boldsymbol{\mu}_0\right)\end{aligned}$$

3 MAP solution for Linear Regression

In class we solved for the maximum likelihood estimator for linear regression with polynomial basis functions. In this exercise you will solve for the *maximum a posterior* (MAP) solution: $\mathbf{w}_{\mathrm{MAP}} = \left(\mathbf{\Phi}^T\mathbf{\Phi} + \lambda\mathbf{I}\right)^{-1}\mathbf{\Phi}^T\mathbf{t}$. For this problem we assume N training vectors $\{\mathbf{x}_n\}_{n=1}^{N}$, each of which is mapped using basis functions to $\boldsymbol{\phi}_n$. In the training set, the data come in input-output pairs, i.e. $\{\mathbf{x}_n, t_n\}$. We assume that one of the basis functions is the constant 1, and there are $M-1$ other basis function in $\boldsymbol{\phi}_n$. We also have the following information:

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

Answer the following:

1. Write down the likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ using a) a product over N and b) in vector/matrix form. Tip: 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.

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

2. Write down the prior $p(\mathbf{w})$ (by expanding the expression for multivariate Gaussian distribution). Compute its log.

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

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

3. Write down an expression for the posterior over $\mathbf{w}$. Remember this will involve applying Bayes rule to the prior, likelihood, and evidence. The evidence will require an integral. You do not need the analytic form for the evidence, but you need the correct variables and conditioning variables, e.g. something like $p(a|b, c)$ where you define a, b, and c.

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

4. Compute the log-posterior, both for the a) and b) likelihood forms from above. Collect everything that does not depend on $\mathbf{w}$ into a constant C. What parts of the previous expression do not depend on $\mathbf{w}$? Why is finding the MAP much simpler than finding the full posterior distribution?

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

5. Solve for $\mathbf{w}_{\mathrm{MAP}}$ by a) taking the derivative of the log-posterior with respect to $\mathbf{w}$, b) setting it to 0, and c) solving for $\mathbf{w}$. Do this for both forms of likelihood.

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

$$\begin{aligned}\frac{\partial \ln p(\mathbf{w}|\mathcal{D})}{\partial \mathbf{w}} &= -\alpha\mathbf{w} - \beta\sum_{n=1}^{N}\left(t_n - \mathbf{w}^T\boldsymbol{\phi}_n\right)\left(-\boldsymbol{\phi}_n\right) = 0 \\ \alpha\mathbf{w} &= \beta\sum_{n=1}^{N}\left(t_n - \mathbf{w}^T\boldsymbol{\phi}_n\right)\boldsymbol{\phi}_n \\ \alpha\mathbf{w} &= \beta\sum_{n=1}^{N}t_n\boldsymbol{\phi}_n - \underbrace{\mathbf{w}^T\boldsymbol{\phi}_n}_{\text{scalar}}\boldsymbol{\phi}_n \\ \alpha\mathbf{w} &= \beta\sum_{n=1}^{N}t_n\boldsymbol{\phi}_n - \boldsymbol{\phi}_n \quad \underbrace{\mathbf{w}^T\boldsymbol{\phi}_n}_{\text{same as } \boldsymbol{\phi}_n^T\mathbf{w}} \\ \left(\alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\right)\mathbf{w} &= \beta\sum_{n=1}^{N}t_n\boldsymbol{\phi}_n \\ \mathbf{w}_{\mathsf{MAP}} &= \left(\alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\right)^{-1}\beta\sum_{n=1}^{N}t_n\boldsymbol{\phi}_n\end{aligned}$$

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

$$\begin{aligned}\frac{\partial \ln p(\mathbf{w}|\mathcal{D})}{\partial \mathbf{w}} &= -\alpha\mathbf{w} - \beta\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} + \beta\mathbf{\Phi}^T\mathbf{t} = 0 \\ \left(\alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}\right)\mathbf{w} &= \beta\mathbf{\Phi}^T\mathbf{t} \\ \mathbf{w}_{\mathsf{MAP}} &= \left(\alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\beta\mathbf{\Phi}^T\mathbf{t} \\ &= \left(\beta\left(\frac{\alpha}{\beta}\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Phi}\right)\right)^{-1}\beta\mathbf{\Phi}^T\mathbf{t} \\ &= \frac{1}{\beta}\left(\lambda\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\beta\mathbf{\Phi}^T\mathbf{t} \\ &= \left(\lambda\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{t}\end{aligned}$$

6. **BONUS** Our prior for $\mathbf{w}$ assumes the same marginal distribution for each entry in $\mathbf{w}$, including that of the first basis function $\boldsymbol{\phi}_0 = 1$. What is the role this basis function? Why should we avoid placing the same penalty/prior for this basis? Rewrite $p(\mathbf{w})$ so that the first basis function has its own prior/penalty.

   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.