HW2: Gaussians and Estimators

Last modified: 2023-02-16 21:49

Status: **RELEASED.**

How to turn in: Submit PDF to https://www.gradescope.com/courses/496674/assignments/2676965/

Jump to: Problem 1 Problem 2 Problem 3

Questions?: Post to the **hw2** topic on the Piazza discussion forums.

Instructions for Preparing your PDF Report

What to turn in: PDF of typeset answers via LaTeX. No handwritten solutions will be accepted, so that grading can be speedy and you get prompt feedback.

Please use provided LaTeX Template: https://github.com/tufts-ml-courses/cs136-23s-assignments/blob/main/unit2_HW/hw2_template.tex

Your PDF should include (in order):

- Cover page with your full name and Collaboration statement
- Problem 1 answer
- Problem 2 answer
- Problem 3 answer

When you turn in the PDF to gradescope, mark each part via the in-browser Gradescope annotation tool)

How to write your solutions

Each step of a mathematical derivation that you turn in should be:

- justified by at least an accompanying short phrase (e.g. "using Bayes rule" or "by the 2nd provided identity")
- legible and easy to follow

Solutions that lack justifications or skip key steps without showing work will receive poor marks.

Problem 1: Estimators of Variance

Consider the following estimator for the *variance* of a univariate Gaussian, given N observed data points $\{x_n\}_{n=1}^N$:

$$\hat{\sigma}^2(x_1, \ldots x_N) = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\text{true}})^2$$

Here, $\mu_{\text{true}} \in \mathbb{R}$ is the *true* mean value (not an estimator).

Note that $\hat{\sigma}$ here is *not* the maximum likelihood estimator of the variance (Eq. 1.56 in PRML textbook). That formula is similar, but uses the ML-estimate of the mean μ^{ML}. Here, we're assuming we know the *true* mean.

1a

Assume that each value x_n is drawn i.i.d from $x_n \sim \mathcal{N}(\mu_{\text{true}}, \sigma^2_{\text{true}})$.

Given this assumed model, compute the expected value of estimator $\hat{\sigma}^2(x_1, \ldots x_N)$.

1b

Using your result in **1a**, explain if the estimator $\hat{\sigma}^2$ is biased or unbiased. Explain why this differs from the biased-ness of the *maximum likelihood* estimator for the variance, using a justification that involves the mathematical definition of each estimator. (*Hint: Why would one be lower than the other?*).

Problem 2: Recognizing Gaussians and Completing the Square

Partial Credit Option: If you wish, you can solve any subproblem for the $M = 1$ univariate case only, for up to 85% credit. If you do this, please write at the top of the relevant subproblem: "I am solving the M=1 case for partial credit".

2a

Suppose you are told that a vector random variable $x \in \mathbb{R}^M$ has the following log PDF function:

$$\log p(x) = \mathrm{c} - \frac{1}{2} x^T A x + b^T x$$

where A is a symmetric positive definite matrix, b is any vector, and c is any scalar constant.

Show that x has a multivariate Gaussian distribution.

Hint: You can solve this by transforming the log PDF above so that has the form:

$$\log p(x) = \mathrm{const} - \frac{1}{2}(x - \mu)^T S (x - \mu)$$

where the constant is with respect to x, and you define a mean vector $\mu \in \mathbb{R}^M$ and precision matrix S (symmetric, positive definite) in terms of A, b above.

Problem 3: Predictive Posteriors for Gaussians

Partial Credit Option: If you wish, you can solve any subproblem for the $M = 1$ univariate case only, for up to 85% credit. If you do this, please write at the top of the relevant subproblem: "I am solving the M=1 case for partial credit".

In this problem, we'll prove that for the Gaussian-Gaussian model of linear regression, each additional data point observed in the training set cannot make the predictive posterior's variance increase. It must be either equal to what it was before, or smaller.

We assume the following probabilistic model for weight vectors $w \in \mathbb{R}^M$ and observed scalar responses $t_n \in \mathbb{R}$.

Prior

$$p(w) = \mathcal{N}(w | 0_M, \alpha^{-1} I_M)$$

where

- $\alpha > 0$ is a known precision hyperparameter
- $0_M \in \mathbb{R}^M$ is the all-zero vector
- I_M is the $M \times M$ identity matrix

Likelihood (conditionally iid)

$$p(t_{1:N} | w) = \prod_{n=1}^{N} \mathcal{N}(t_n | w^T \phi(x_n), \beta^{-1})$$

where

- $\beta > 0$ is a known precision hyperparameter
- $w \in \mathbb{R}^M$ is weight vector (modeled as a random variable with prior above)
- $\phi(x_n) \in \mathbb{R}^M$ is a known feature vector for n-th train example

Here, we've assumed conditional independence in the likelihood. If w is known, each t_n only depends on w, and also knowing other t_i values does not improve our predictions of t_n further. We'll also assume for new test points that t_* is independent of other t_i given w.

Posterior predictive for N train examples

This model has the predictive distribution given in the textbook (Eq. 3.58):

$$\begin{aligned}
p(t_* | t_{1:N}) = \int_w p(t_*, w | t_{1:N}) dw &= \mathcal{N}(t_* | m_N^T \phi(x_*), \sigma_N(x_*)^2) \\
\sigma_N^2(x_*) &= \beta^{-1} + \phi(x_*)^T S_N \phi(x_*) \\
S_N^{-1} &= \alpha I_M + \beta \Phi_{1:N}^T \Phi_{1:N} && \text{see Eq. 3.51} \\
m_N &= \beta S_N \Phi_{1:N}^T t_{1:N} && \text{see Eq. 3.50}
\end{aligned}$$

Here, we have defined:

- $\Phi_{1:N}$ is the $N \times M$ feature matrix obtained by stacking the feature vectors $\phi(x_n)$ for each example n in train set of size N.
- $t_{1:N}$ is the $N \times 1$ column vector obtained by stacking the responses t_n for each example n in train set of size N
- m_N is the M-length vector defining the mean of the Gaussian posterior for w given train set of size N
- S_N is the $M \times M$ covariance matrix parameter of the Gaussian posterior for w given train set of size N
- $\sigma_N^2(x_*) > 0$ is the variance of the posterior predictive at a test example whose input feature is x_*

From N to $N+1$ train examples

Imagine adding one extra feature-response pair x_{N+1}, t_{N+1} to the training set. After this addition, we can define $m_{N+1}, S_{N+1}, \sigma_{N+1}$ by conditioning on the expanded train set of $N+1$ examples.

3a

Show that we can write $S_{N+1}^{-1} = S_N^{-1} + vv^T$ for some vector $v \in \mathbb{R}^M$.

3b

Next, consider the following identity, which holds for any invertible matrix A:

$$(A + vv^T)^{-1} = A^{-1} - \frac{(A^{-1}v)(v^T A^{-1})}{1 + v^T A^{-1} v}$$

Substitute $A = S_N^{-1}$ and v as defined in **3a** into the above. Simplify to write an expression for S_{N+1} in terms of S_N.

3c

Show that $\sigma_{N+1}^2(x_*) - \sigma_N^2(x_*) = \phi(x_*)^T \left[S_{N+1} - S_N\right] \phi(x_*)$

3d

Finally, plug your result from **3b** defining S_{N+1} into **3c**, plus the fact that S_N must be positive definite, to show that:

$$\sigma_{N+1}^2(x_*) \leq \sigma_N^2(x_*)$$

This would prove that the predictive variance *cannot increase* with each additional data point. In other words, we will never be "less certain" about a prediction we make if we gather more data.
