INSTRUCTOR SOLUTION for HW2

Collaboration Statement:

Total hours spent: 3 hour

We consulted the following resources:

- Bishop's PRML textbook
- Murphy's 2012 textbook

Links: [HW2 instructions] [collab. policy]

Contents

1a: Problem Statement

Compute the expected value of estimator $\hat{\sigma}^2(x_1, \ldots x_N)$, where

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

1a: Solution

The desired expectation is:

$$\mathbb{E}_{x_n \sim \mathcal{N}(\mu_{\text{true}}, \sigma^2_{\text{true}})} \left[\hat{\sigma}^2(x_1, \ldots x_N) \right] = \mathbb{E}_{x_n \sim \mathcal{N}(\mu_{\text{true}}, \sigma^2_{\text{true}})} \left[\frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{\text{true}})^2 \right] \tag{2}$$

Expanding the quadratic, we obtain:

$$= \mathbb{E}_{x_n \sim \mathcal{N}(\mu_{\text{true}}, \sigma^2_{\text{true}})} \left[\frac{1}{N} \sum_{n=1}^{N} (x_n^2 - 2x_n \mu_{\text{true}} + \mu^2_{\text{true}}) \right] \tag{3}$$

Applying linearity of expectations to bring the expectation inside the sum, we have:

$$= \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[x_n^2] - \mathbb{E}[x_n \mu_{\text{true}}] + \mathbb{E}[\mu^2_{\text{true}}] \tag{4}$$

Because μ_{true} is a *constant* wrt our random variables x, so we can simplify to:

$$= \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[x_n^2] - \mathbb{E}[x_n] \mu_{\text{true}} + \mu^2_{\text{true}} \tag{5}$$

Next, we use two facts about Gaussian random variables: $\mathbb{E}[x_n] = \mu_{\text{true}}$ and $\mathbb{E}[x_n^2] = \mu^2_{\text{true}} + \sigma^2_{\text{true}}$. Substituting these in, we have

$$= \frac{1}{N} \sum_{n=1}^{N} \mu^2_{\text{true}} + \sigma^2_{\text{true}} - 2\mu^2_{\text{true}} + \mu^2_{\text{true}} \tag{6}$$

Simplifying, all μ_{true} terms cancel and we find that the expected value of the estimator is the true variance:

$$= \frac{1}{N} \sum_{n=1}^{N} \sigma^2_{\text{true}} = \sigma^2_{\text{true}} \tag{7}$$

1b: Problem Statement

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?).

1b: Solution

This estimator is *unbiased*. Its expected value of this estimate of the variance parameter is equal to the true parameter we are trying to estimate.

In contrast, consider the ML estimator of variance:

$$\sigma^2_{ML}(x_1, \dots x_N) = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{\text{ML}})^2 \tag{8}$$

As shown in the textbook (and in the lecture notes), this estimator is *biased*: its expected value is $\mathbb{E}[\sigma^2_{\text{ML}}] = \frac{N-1}{N} \sigma^2_{\text{true}}$, which is slightly smaller than the true variance: fracN-1N $\sigma^2_{\text{true}} < \sigma^2_{\text{true}}$

Comparing the two estimators, both compute the sum of squared errors between a mean value μ and each training point x_n. The only difference is whether we use μ_{true} or μ_{ML}. The estimator μ_{ML} is chosen to *maximize likelihood* (equivalently, minimize sum-of-squared-errors) for the given dataset of size N. Thus we can always be sure that μ_{ML} leads to a smaller sum-of-squared errors (otherwise by counterexample μ_{true} would be a better ML estimator for μ). So we know:

$$\sum_{n=1}^{N} (x_n - \mu_{\text{ML}})^2 \leq \sum_{n=1}^{N} (x_n - \mu_{\text{true}})^2 \tag{9}$$

Based on this analysis, it makes sense that the ML-estimated variance is typically underestimated: $\sigma^2_{\text{ML}} \leq \sigma^2_{\text{true}}$, and really cannot be overestimated. Thus, using the ML-estimate of the mean leads to bias in the ML-estimate of the variance.

2a: Problem Statement

Suppose vector r.v. $x \in \mathbb{R}^M$ has the following log PDF function:

$$\log p(x) = \text{c} - \frac{1}{2} x^T A x + b^T x \tag{10}$$

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.

2a: Solution

Strategy: transform a known Gaussian PDF into a similar form as above.

Suppose we have a Gaussian random variable with precision matrix S (symmetric, positive definite), and mean vector μ. We could rewrite the log PDF as:

$$\begin{aligned}
\log p(x) &= \text{const} - \frac{1}{2}(x-\mu)^T S (x-\mu) && \qquad (11)\\
&= \text{const} - \frac{1}{2}\left(x^T S x - \mu^T S x - x^T S \mu + \mu^T S \mu\right) && \text{By expanding the quadratic}\\
&= \text{const} - \frac{1}{2}\left(x^T S x - 2\mu^T S x + \mu^T S \mu\right) && \text{By symmetry of } S.\\
&= \text{const} - \frac{1}{2}\left(x^T S x - 2\mu^T S x\right) && \text{Gather } \mu^T S \mu \text{ into constant}\\
&= \text{const} - \frac{1}{2} x^T S x - (S\mu)^T x && \text{Simplifying algebra}
\end{aligned}$$

where $(S\mu)^T = \mu^T S$ by definition of transpose of product when S is symmetric.

Now, let us define two new symbols (remember to read $\triangleq$ as "is defined as")

$$\begin{aligned}
A &\triangleq S, && A \text{ is an } M \times M \text{ symmetric, positive definite matrix} && (12)\\
b &\triangleq S\mu, && b \text{ is a } M \times 1 \text{ column vector}
\end{aligned}$$

Using these symbols, we can re-write our last line above as

$$\log p(x) = \text{const} - \frac{1}{2} x^T A x - b^T x$$

and we have arrived at our desired result.

BONUS: we can write S, μ in terms of A, b:

$$S = A, \qquad \mu = S^{-1} b = A^{-1} b \tag{13}$$

3a: Problem Statement

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

3a: Solution

After observing N examples, we can write S_N as

$$\begin{aligned} S_N^{-1} &= S_0^{-1} + \beta \Phi_{1:N}^T \Phi_{1:N} && \text{by the definition of } S_N \quad (14) \\ &= S_0^{-1} + \beta \sum_{n=1}^{N} \underbrace{\phi(x_n)\phi(x_n)^T}_{\text{outer product, shape } M \times M} \end{aligned}$$

using the fact that matrix $\Phi_{1:N}$ is made up by stacking up feature vectors $\phi(x_n) \in \mathbb{R}^M$ as rows, and using the view of a matrix multiply as a sum of outer products.

Similarly, after observing $N + 1$ examples (one more than above) we can write:

$$S_{N+1}^{-1} = S_0^{-1} + \beta \sum_{n=1}^{N+1} \phi(x_n)\phi(x_n)^T \tag{15}$$

Rewriting the sum over $N + 1$ examples into two terms, a sum over the first N examples and a separate last example, we have:

$$S_{N+1}^{-1} = \underbrace{S_0^{-1} + \beta \sum_{n=1}^{N} \phi(x_n)\phi(x_n)^T}_{S_N^{-1}} + \beta \phi(x_{N+1})\phi(x_{N+1})^T \tag{16}$$

Splitting $\beta = \sqrt{\beta}\sqrt{\beta}$, we can rewrite this in the desired form of the S_N term plus an outer product of a vector v:

$$\begin{aligned} S_{N+1}^{-1} &= S_N^{-1} + \left(\sqrt{\beta}\phi(x_{N+1})\right)\left(\sqrt{\beta}\phi(x_{N+1})\right)^T && (17) \\ &= S_N^{-1} + vv^T, \quad v \triangleq \sqrt{\beta}\phi(x_{N+1}) && (18) \end{aligned}$$

We've now defined S_{N+1}^{-1} in terms of S_N^{-1} and an M-dimensional vector v, as desired.

3b: Problem Statement

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

$$(A + vv^T)^{-1} = A^{-1} - \frac{(A^{-1}v)(v^T A^{-1})}{1 + v^T A^{-1} v} \tag{19}$$

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.

3b: Solution

Substituting $A = S_N^{-1}$ (and thus $A^{-1} = S_N$), we have

$$(S_N^{-1} + vv^T)^{-1} = S_N - \frac{(S_N v)(v^T S_N)}{1 + v^T S_N v} \tag{20}$$

Recalling that $S_{N+1}^{-1} = S_N^{-1} + vv^T$, we rewrite the left-hand side as:

$$S_{N+1} = S_N - \frac{1}{1 + v^T S_N v}(S_N v)(v^T S_N) \tag{21}$$

We've now defined S_{N+1} in terms of S_N (and v), as desired.

3c: Problem Statement

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

3c: Solution

We start by restating the general definition of the predictive variance after seeing training sets of size N and $N+1$ examples:

$$\begin{aligned} \sigma_N^2(x_*) &= \beta^{-1} + \phi(x_*)^T S_N \phi(x_*) \\ \sigma_{N+1}^2(x_*) &= \beta^{-1} + \phi(x_*)^T S_{N+1} \phi(x_*) \end{aligned} \tag{22}$$

Taking the difference (second line minus first line), the β terms cancel, and we have

$$\sigma_{N+1}^2(x_*) - \sigma_N^2(x_*) = \phi(x_*)^T (S_{N+1} - S_N) \phi(x_*) \tag{23}$$

which achieves our goal.

3d: Problem Statement

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:

$$\sigma^2_{N+1}(x_*) \leq \sigma^2_N(x_*) \tag{24}$$

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

3d: Solution

From 3b, we know $S_{N+1} - S_N = \frac{-1}{1+v^T S_N v}(S_N v)(v^T S_N)$. Plugging into 3c gives

$$\sigma^2_{N+1}(x_*) - \sigma^2_N(x_*) = \frac{-1}{1 + v^T S_N v} \cdot \phi(x_*)^T \left[(S_N v)(v^T S_N)\right] \phi(x_*) \tag{25}$$

where we've simplified by bringing the scalar term out front. Second, using the associativity of matrix-vector multiplication, we can regroup the multiplies as:

$$\sigma^2_{N+1}(x_*) - \sigma^2_N(x_*) = \frac{-1}{1 + v^T S_N v} \left(\phi(x_*)^T S_N v\right) \left(v^T S_N \phi(x_*)\right) \tag{26}$$

Because S_N is symmetric, we know scalar $a^T S_N b = b^T S_N a$ for any vectors a and b, and thus our difference of variances becomes a product of two scalars:

$$\begin{aligned}\sigma^2_{N+1}(x_*) - \sigma^2_N(x_*) &= \frac{-1}{1 + v^T S_N v} \left(\phi(x_*)^T S_N v\right) \left(\phi(x_*)^T S_N v\right) \\ &= \underbrace{\frac{-1}{1 + v^T S_N v}}_{\text{always}<0} \underbrace{\left(\phi(x_*)^T S_N v\right)^2}_{\text{always}\geq 0} \\ &\leq 0\end{aligned} \tag{27}$$

The first scalar is always *negative*, because the numerator is negative and the denominator is positive (at least 1). Recall that S_N is positive definite, thus by definition, $v^T S_N v \geq 0$ for any non-zero vector v.

The second scalar is *always non-negative*, because it is a square of the scalar $(\phi(x_*)^T S_N v)$ and squares are never negative.

Thus, together, the product of the first (negative) and second (non-negative) will be *non-positive*. This implies σ^2_{N+1} is always less than or equal to σ^2_N, as desired.