Quiz 3 Solutions

02 Bias-Variance Decomposition

by Alvin Wan . alvinwan.com/cs189/fa17

Treat this as an exam situation. You will be given 5 minutes to complete this quiz.

1 Just Bias and Variance

Let us consider a probabilistic perspective, where the data is now "random". We believe that our data is sampled from a true distribution, and our goal is to uncover that underlying distribution. Take your data to be i.i.d. $\{\vec{x}_i\}_{i=1}^n$ where $\vec{x}_i \sim \mathcal{N}(\mu, \sigma^2 I), \vec{x}_i \in \mathbb{R}^d$.

1. Say you have only one point (e.g., $n = 1$). Compute the maximum likelihood estimate $\hat{\mu}$ for $E[X]$. What is $\hat{\mu}$?

   Solution: It's simply your only point. The MLE is

   $$\hat{\mu} = \sum_{i=1}^{n} x_i = x$$

2. Compute the mean-squared error (MSE), $E\|\hat{\mu} - \mu\|_2^2$. Express it terms of μ, σ, d.

   Solution: We have two available methods, one vectorized approach and one component-wise.

   Solution 1 Break down the vector component-wise.

   $$\begin{aligned} E\|\hat{\mu} - \mu\|_2^2 &= E[\|x - \mu\|_2^2] \\ &= \sum_i E[(x_i - \mu_i)^2] \\ &= d\sigma^2 && \text{definition of variance} \end{aligned}$$

   Solution 2 Instead of breaking this down component-wise, we can also realize this by applying the trace.

$$
\begin{aligned}
E\|\hat{\mu} - \mu\|_2^2 &= E[\|x - \mu\|_2^2] \\
&= E[(x-\mu)^T(x-\mu)] \\
&= E[Tr((x-\mu)^T(x-\mu))] && \text{Trace of a scalar is itself} \\
&= E[Tr((x-\mu)(x-\mu)^T)] && \text{Cyclic property of trace} \\
&= E[Tr(\Sigma)] && \text{where } \Sigma \text{ is the covariance matrix} \\
&= d\sigma^2
\end{aligned}
$$

3. Instead of MLE, say we develop an affine model to estimate μ, $\hat{\mu}_2 = \alpha x + \beta$. What is $E[\hat{\mu}_2]$?

   Solution:

$$E[\hat{\mu}_2] = E[\alpha x + \beta] = \alpha E[x] + \beta = \alpha\mu + \beta$$

4. For simplicity, say $\beta = 0$. Compute the MSE for $\hat{\mu}_2$.

   Solution: We have two available methods, one vectorized approach and one component-wise.

   Solution 1 Expand the terms out.

$$
\begin{aligned}
E[\|\hat{\mu_2} - \mu\|_2^2] &= E[\| \underbrace{\alpha x}_{a} - \underbrace{\mu}_{b} \|_2^2] \\
&= \underbrace{\alpha^2 E[\|x\|_2^2]}_{a^2} - \underbrace{2\alpha E[x\mu]}_{2ab} + \underbrace{E[\|\mu\|_2^2]}_{b^2} \\
&\overset{(a)}{=} \alpha^2 d\sigma^2 + \alpha^2\|\mu\|_2^2 - 2\alpha\|\mu\|_2^2 + \|\mu\|_2^2 \quad \mu \text{ is constant w.r.t. } x \\
&= \alpha^2 d\sigma^2 + (\alpha-1)^2\|\mu\|_2^2
\end{aligned}
$$

   (a) We can use the fact that $\text{var}(x_i) = E[x_i^2] - E[x_i]^2$. Rearrange to get $E[x_i^2] = \text{var}(x_i) + E[x_i]^2$.

$$E[\|x\|_2^2] = \sum_i E[x_i^2] = \sum_i (var(x_i) - E[x_i]^2) = \sum_i (\sigma^2 - \|\mu\|^2) = \|\mu\|^2 + d\sigma^2$$

   Solution 2 Alternatively, use the trace and add 0.

$$
\begin{aligned}
&E[\|\hat{\mu}_2-\mu\|_2^2] \\
&= E[\|\hat{\mu}_2 - E[\hat{\mu}_2] + E[\hat{\mu}_2] - \mu\|_2^2] \\
&= E[\|\underbrace{\hat{\mu}_2-\alpha\mu}_{a} + \underbrace{\alpha\mu-\mu}_{b}\|_2^2] && \text{Recall } \beta = 0 \\
&= E[\|\underbrace{\hat{\mu}_2-\alpha\mu\|_2^2}_{a^2}] + E[\|\underbrace{2(\hat{\mu}_2-\alpha\mu)^T(\alpha\mu-\mu)]}_{2ab} + E[\underbrace{\|\alpha\mu-\mu\|_2^2}_{b^2}] \\
&\overset{(a)}{=} \alpha^2 d\sigma^2 + E[\|2(\hat{\mu}_2-\alpha\mu)^T(\alpha\mu-\mu)] + E[\|\alpha\mu-\mu\|_2^2] \\
&\overset{(b)}{=} \alpha^2 d\sigma^2 + E[\|\alpha\mu-\mu\|_2^2] \\
&= \alpha^2 d\sigma^2 + (\alpha-1)^2\|\mu\|_2^2 && \text{All constant w.r.t. x}
\end{aligned}
$$

(a) Recall that $\hat{\mu}_2 = \alpha x + \beta = \alpha x$, so

$$E[\|\hat{\mu}_2-\alpha\mu\|_2^2] = E[\|\alpha(x-\mu)\|_2^2] = \alpha^2 E[\|x-\mu\|_2^2]$$

, where the expectation is $d\sigma^2$ per part 1, to get $\alpha^2 d\sigma^2$. (b) Rewrite the term, plugging in $\hat{\mu}_2$.

$$E[2(\hat{\mu}_2-\alpha\mu)^T(\alpha\mu-\mu)] = E[2(\alpha(x-\mu))^T(\alpha\mu-\mu)]$$

Note that all terms except x are constant and that $E[x-\mu] = E[x] - \mu = 0$, since $E[x] = \mu$ by construction. Thus, the second term is 0.