Machine Learning and Bayesian Inference

Dr Sean Holden

Computer Laboratory, Room FC06

Telephone extension 63725

Email: sbh11@cam.ac.uk

https://www.cl.cam.ac.uk/~sbh11

Question and Answer Session 2

Question:

"I have a question on the notation on two different slides from the MLBI course: slide 46 and slide 77. On slide 46, we need to minimise the equation, which contains the underlying variance σ *of the distribution as a parameter. On slide 77, this variance* σ *does not appear anymore. Something similar happens between slides 49 and 81, on the MAP algorithm.*

While I know that it does not make a difference for maximum likelihood estimation, should not the variance make a difference for MAP? Since the variance λ *of the prior distribution of* $\mathbf{w}$ *is kept in the equation, should not* σ *be kept as well?"*

Question and Answer Session 2

Slide 46:

$$\mathbf{w}_{\text{opt}} = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - h_{\mathbf{w}}(\mathbf{x}_i))^2$$

Slide 77:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} (y_i - h_{\mathbf{w}}(\mathbf{x}_i))^2.$$

The questioner is *entirely correct*: the outcome of the minimization doesn't depend on $1/\sigma^2$ because it's just a constant factor.

Slide 49:

$$\mathbf{w}_{\text{opt}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left[\frac{1}{2\sigma^2} \sum_{i=1}^{m} \left((y_i - h_{\mathbf{w}}(\mathbf{x}_i))^2 \right) + \frac{\lambda}{2} ||\mathbf{w}||^2 \right].$$

Slide 81:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} \left((y_i - h_{\mathbf{w}}(\mathbf{x}_i))^2 \right) + \frac{\lambda}{2} ||\mathbf{w}||^2.$$

The questioner is again *entirely correct* to note that perhaps σ and λ should now be treated separately.

Later on, when looking at the Bayesian formulation in more detail—slide 178 onward—we shall see that this becomes important. (They are rolled into the *hyperparameters* α and β.) *However...*

…*in practice*, when implementing MAP rather than the full Bayesian solution, we can effectively roll both σ and λ into a single *regularization parameter* λ'.

$$\begin{aligned}
\mathbf{w}_{\text{opt}} &= \operatorname*{argmin}_{\mathbf{w}} \left[\frac{1}{2\sigma^2} \sum_{i=1}^{m} \left((y_i - h_{\mathbf{w}}(\mathbf{x}_i))^2 \right) + \frac{\lambda}{2} ||\mathbf{w}||^2 \right] \\
&= \operatorname*{argmin}_{\mathbf{w}} \sigma^2 \left[\cdots \right] \\
&= \operatorname*{argmin}_{\mathbf{w}} \left[\frac{1}{2} \sum_{i=1}^{m} \left((y_i - h_{\mathbf{w}}(\mathbf{x}_i))^2 \right) + \frac{\lambda'}{2} ||\mathbf{w}||^2 \right]
\end{aligned}$$

where

$$\lambda' = \sigma^2 \lambda.$$

The single parameter λ' can then be optimized using *cross-validation* for example.

Question:

"Sometimes I get a bit confused on hypothesis tests, and on what exactly can and can't be tested for. How would you test for two methods giving equivalent performance?"

Question and Answer Session 2

"Sometimes I get a bit confused on hypothesis tests, and on what exactly can and can't be tested for."

This is not surprising:

- Statistical testing is a *MASSIVE* subject.
- To be anything like comprehensive, I'd need to use *all sixteen lectures*.
- This course covers the *bare minimum*.

What's important is the *take-home message*: if you want to be taken seriously, then apply an *appropriate test of significance* and report the result.

What constitutes an appropriate test will depend on the circumstances, and may not be covered here: Stuart-Maxwell test, Kolmogorov-Smirnov, Welch's t-test, Analysis of Variance, Mann-Whitney U test, McNemar's test, Wilcoxon signed-rank test, Bayesian alternatives...

Question and Answer Session 2

As far as this course is concerned:

- **Slide 168:** Confidence interval if I *estimate* a *mean* using m samples.
- **Slide 172:** Confidence interval if I estimate the *difference* in *accuracy* between two *already trained* classifiers using m samples.
- **Slide 177:** Confidence interval if I estimate the expected difference in performance between two *algorithms* when training on sets of m examples.

Question and Answer Session 2

"How would you test for two methods giving equivalent performance?"

Answer: I wouldn't!

- In machine learning, it is essentially *unheard of* to test for two methods having *the same performance*.
- One tests to establish some level of confidence that performance is *improved*.