Bayesian Machine Learning – Review

I have annotated the document where I have spotted typos or formalities that are wrong. Apart from that, I still have some general remarks:

In the first section, you derive Bayes' rule. I think it would be a good idea to briefly mention why it is important (i.e. you can compute $P(X|Y)$ with $P(Y|X)$) since it is otherwise unclear why one would ever use it.

In the explanation of the number game, it is a bit unclear what you mean with "positive examples". You should maybe explain that further. Also, when discussing the likelihood, you state that "the likelihood can be deterministically computed [...] using the extension of a concept, which is defined as all elements, that belong to it". In my opinion this wording is somewhat ambiguous and I have only understood what you mean with the extension of a concept after reading Murphy myself, so I suggest you rephrase this.

After you present the posterior, you mention how it relates to the MAP and MLE estimate. However, you only introduce those concepts afterwards, which is one reason why that part is difficult to understand. Perhaps it would be better if you first explain MLE and MAP and afterwards outline the relation to the posterior.

Why do you define the MLE using argmin? All formulations I have seen so far, including Mitchell and Murphy, define it with argmax, which also makes intuitive sense. If this is not an error, you should definitely explain where the argmin comes from (maybe you confused the likelihood with the negative likelihood?).

A natural question that your paper doesn't answer is when one would choose the MLE vs the MAP estimation.

In the definition of Bayesian Networks it is kind of missing what the nodes and edges of the DAG represent. Of course, this becomes somewhat clear in the examples, but a short explanation would be nice.

In the section on linear models, you use the notation x^n to denote an input vector. I suggest using a slightly different notation such as $x^{(n)}$ or x_n as one can otherwise mistake it with the operation "to the power of n".

In my opinion, you should elaborate further on why methods such as ridge regression produce better results. From your explanation alone it is not clear why adding such a regularization term might benefit the model.

I have to admit that I couldn't quite follow your explanation of Gaussian processes. Of course, this is a rather complicated topic, but maybe a small example or slightly different explanation could help.