Learning the Prior from Data?

• How do we tune the hyper-parameters in Bayesian methods?

• Adapting our usual validation set approach:
  – Split into a training and validation set.
  – For different hyper-parameter values:
    • Compute some measure of “test error”.
      – For density estimation, this could be the posterior predictive for the validation set given the training set.
      – For supervised learning, you could make predictions on the validation set and measure validation set error.
    – Choose the hyper-parameters with the highest value.

• Advantage:
  – Directly tunes hyper-parameters to achieve good performance on new data.

• Disadvantage:
  – Optimization bias: can start to overfit to the validation set.
  – Slow! If you try 10 values for ‘k’ hyper-parameters, there are $10^k$ values to try.
Learning the Prior from Data?

• **Empirical Bayes:**
  – Optimize the likelihood of the data given the hyper-parameters.

\[
\hat{\alpha} \in \arg \max_{\alpha} \mathbb{E}[p(X | \alpha)] = \arg \max_{\alpha} \int p(X | \alpha) p(\alpha | \Theta) \, d\alpha
\]

  • This is called the “marginal likelihood” or the “evidence” function.
  – It can be computed by marginalizing over parameters.
  – It is the denominator we ignore when we do MAP estimation: \( p(\Theta | X) = \frac{p(X | \Theta)p(\Theta | A)}{p(X | A)} \).
  • Empirical Bayes is also called “type II maximum likelihood” or “evidence maximization”.
    – This is doing **MLE for the hyper-parameters**.

• **Advantage:**
  – **Fast**! Might have a closed-form solution or allow using gradient descent (assuming conjugate prior).

• **Disadvantage:**
  – It is **not directly testing** the performance on new data.
  – Optimization bias: can start to **overfit the marginal likelihood** (could increase/decrease test performance).
Marginal Likelihood with Conjugate Priors

- **Marginal likelihood** has **closed-form** when using conjugate priors.
  - It is proportional to ratio of posterior/prior normalizing constants.

- We will show this for the Bernoulli-Beta model:

\[
\begin{align*}
\rho(X | \theta) &= \theta^{n}(1-\theta)^{n_0} \\
\rho(\theta | \alpha, \beta) &= \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{Z(\alpha, \beta)} \\
\rho(\theta | X, \alpha, \beta) &= \frac{\Theta^{(n+\alpha)-1}(1-\theta)^{(n_0+\beta)-1}}{Z(n+\alpha, n_0+\beta)} \\

Z(\alpha, \beta) &= \int \Theta^{\alpha-1}(1-\Theta)^{\beta-1} d\Theta \\
Z(n+\alpha, n_0+\beta) &= \int \Theta^{(n+\alpha)-1}(1-\Theta)^{(n_0+\beta)-1} d\Theta \\
\end{align*}
\]
Marginal Likelihood with Conjugate Priors

• For the Bernoulli-beta model we have marginal likelihood of:

\[ p(Y | \alpha, \beta) = \frac{Z(n + \alpha, n_0 + \beta)}{Z(\alpha, \beta)} \]

– For other distributions the ratio might be multiplied by a constant.
  • By similar argument, posterior predictive for new data with counts \( \tilde{n}_1 \) and \( \tilde{n}_0 \) is:

\[ \frac{Z(n + \tilde{n}_1 + \alpha, n_0 + \tilde{n}_0 + \beta)}{Z(n + \alpha, n_0 + \beta)} \]

• Empirical Bayes maximizes marginal likelihood in terms of \( \alpha \) and \( \beta \).
  – More useful when we have many hyper-parameters.
  – Could be used for categorical-Dirichlet model’s ‘k’ hyper-parameters.
  – In some cases is equivalent to leave-one-out cross-validation.
    • The most-extreme form of cross-validation (in a good way).
Learning Principles for Predicting “0 or 1 Next?”

• Maximum likelihood:
  \[ \hat{\theta} \in \arg\max_{\theta} \mathbb{E}[\rho(X|\theta)] \quad \hat{x} \in \arg\max_{x} \mathbb{E}[\rho(x|\theta)] \]

• MAP:
  \[ \hat{\theta} \in \arg\max_{\theta} \mathbb{E}[\rho(\theta|X,\alpha,\beta)] \quad \hat{x} \in \arg\max_{x} \mathbb{E}[\rho(x|\theta)] \]

• Bayesian (no “learning”):
  \[ \hat{x} \in \arg\max_{x} \mathbb{E}[\rho(x|X,\alpha,\beta)] \equiv \arg\max_{x} \mathbb{E} \left[ \int \rho(\theta|X,\alpha,\beta) \rho(x|\theta) d\theta \right] \]

• Empirical Bayes:
  \[ \hat{\alpha}, \hat{\beta} \in \arg\max_{\alpha,\beta} \mathbb{E}[\rho(X|\alpha,\beta)] \quad \hat{\gamma} \in \arg\max_{\gamma} \mathbb{E}[\rho(x|X,\hat{\alpha},\hat{\beta})] \]
Bayesian Hierarchy

• Maximum likelihood estimation can do weird things.
  – Predict zero probability for events not seen in training.
  – Pick a highly-unlikely model that exactly fits the training data.

• MAP estimation improves MLE by adding a prior on the parameters.
  – But by only using one parameter estimate this leads to sub-optimal decisions.

• Bayesian inference over parameters makes optimal decisions.
  – Avoids overfitting, and decisions follow rules of probability.
    • No optimization bias because no optimization.
  – But this relies on have a good choice of prior/hyper-parameters.

• Empirical Bayes uses data to find a good prior.
  – Tends to be less sensitive to overfitting than regular MLE.
  – But has an optimization bias: can still overfit the hyper-parameters.
  – In my experience, more likely to “just be weird” than actual overfitting.
Bayesian Hierarchy

• To fix empirical Bayes issues:
  – We can put a prior on the hyper-parameters.
  – Sometimes called a “hyper-prior”, that has “hyper-hyper-parameters”.
    • Seriously!
  – But by only using one parameter estimate this leads to sub-optimal decisions.

• So use Bayesian inference over parameters and hyper-parameters:
  – You would integrate over all values of the parameters and hyper-parameters.
    • Unfortunately, we often do not have a “conjugate hyper-prior” for the prior.
  – This avoids overfitting, but now we rely on having a good choice of hyper-prior.

• And then could consider empirical Bayes over hyper-hyper-parameters...
  – This was one the hottest ML topics before deep learning came back.
Next Topic: Hierarchical Bayes
Motivating Example: Medical Treatment

• Consider modeling \textit{probability that a medical treatment will work.}
  – But this probability \textit{depends on the hospital} where treatment is given.
• So we have binary examples \(x^1, x^2,...,x^n\).
  – We also have a number \(z^i\) saying “what population it came from”.
    • This is a common \textit{non-IID} setting: examples are \textit{only IID} within each group.

\[
\begin{array}{c|c|c}
\text{Worked?} & \text{Hospital} \\
1 & 1 \\
0 & 4 \\
0 & 3 \\
1 & 2 \\
0 & 3 \\
\end{array}
\]

\(X = \begin{pmatrix}
1 \\
0 \\
0 \\
1 \\
0 \\
\end{pmatrix}\)

\(p(x = 1 \mid z=2) = 0.4\)

• Other examples:
  – “What are the covid proportions for different cities?”
  – “Which of my stores will sell over 100 units of product?”
  – “What proportion of users will click my adds on different websites?”
Independent Model for Each Group

• We could consider a simple independent model for each group:
  – Use a parameter $\theta_j$ for each hospital ‘j’.
    \[ x_i | z_i \sim \text{Ber}(\theta_{z_i}) \]
  – Fit each $\theta_j$ using only the data from hospital ‘j’.
    • If we have ‘k’ hospitals, we solve ‘k’ IID learning problems.

• Problem: we may not have a lot of data for each hospital.
  – Can we use data from a hospital with a lot of data to learn about others?
  – Can we use data across many hospitals to learn with less data?
  – Can we say anything about a hospital with no data?
Dependencies from Using a Common Prior

• Common approach: assume the $\theta_j$ are drawn from a common prior.

$$x^i | z^i \sim \text{Ber}(\theta^i) \quad \theta_j \sim \text{Beta}(\alpha, \beta)$$

• This introduces a dependency between the $\theta_j$ values.
  – For example, if $\alpha = 5$ and $\beta = 2$:
    • This is like we imagine seeing 5 extra “success” and 2 “failures” at each hospital.

• In this setting the $\theta_j$ are conditionally independent given $\alpha$ and $\beta$.
  – With a fixed prior, we cannot learn about one $\theta_j$ using data from another.
    • So for a new hospital, the posterior over $\theta_j$ is the prior.

• In this setting, we want to learn the hyper-parameters.
Hierarchical Bayesian Modeling

• Consider using a **hyper-prior**:

\[
\begin{align*}
  x \mid z_i & \sim \text{Ber}(\theta_{z_i}) \\
  \theta_j & \sim \text{Beta}(\alpha, \beta) \\
  \alpha, \beta & \sim D(\rho, \eta, \omega)
\end{align*}
\]

  – Treating hyper-parameters as random variables, can **learn across groups**.

• With **empirical Bayes** we get fixed estimates of $\tilde{\alpha}$ and $\tilde{\beta}$.
  – Learned prior gives **better estimates of $\theta_j$** for groups with few examples.
  – For a **new hospital**, posterior would default to the learned prior.

• With **hierarchical Bayes** we would integrate over the $\theta_j$'s, $\alpha$, and $\beta$.
  – “Very Bayesian” to handle the unknown parameters/hyper-parameters.
  – Hierarchical models almost always need approximations like Monte Carlo.
Discussion of Hierarchical Bayes

• Many practitioners really like Bayesian models.
  – “Gosh darn, I love Bayesian ensemble methods!”
    • From a domain expert I was collaborating with.
  – Domain expertise can be incorporated into the design of [hyper-]priors.
  – Can model various ways your data may not be IID.
  – We will see some more Bayes tricks.

• Advantage is the nice mathematically framework:
  – Write out all your prior knowledge of relationships between variables.
  – Integrate over variables you do not know.

• Disadvantages:
  – It can be hard to exactly encode your prior beliefs.
  – The integrals get ugly very quickly (there is no “automatic integration”).
Summary

- **Marginal likelihood:**
  - Probability of data given hyper-parameters (integrating over parameters).

- **Empirical Bayes** ("type II MLE" or "evidence maximization").
  - Tune hyper-parameters by optimizing marginal likelihood.
  - Can be used to cheaply tune a huge number of hyper-parameters.
    - If you can efficiently do/approximate the integrals.

- **Hyper-priors:**
  - Putting a prior on the prior.
  - Often needed to make empirical Bayes work, or in hierarchical Bayes.

- **Hierarchical Bayes:**
  - Building models with multiple levels of priors.
  - Often allows learning in non-standard scenarios.
    - We considered the case of non-IID grouped data.

- Next Time: everyone’s favourite loss to take the gradient of.