CPSC 540: Machine Learning

Conjugate Priors

Mark Schmidt

University of British Columbia

Winter 2020

Last Time: Bayesian Predictions and Empirical Bayes

- We've discussed making predictions using posterior predictive,

$$\hat{y} \in \underset{\tilde{y}}{\text{argmax}} \int_w p(\tilde{y} \mid \tilde{x}, w) p(w \mid X, y, \lambda) dw,$$

  which gives optimal predictions given your assumptions.
- We considered empirical Bayes (type II MLE),

$$\hat{\lambda} \in \underset{\lambda}{\text{argmax}}\, p(y \mid X, \lambda), \quad \text{where} \quad p(y \mid X, \lambda) = \int_w p(y \mid X, w) p(w \mid \lambda) dw,$$

  where we optimize marginal likelihood to select model and/or hyper-parameters.
  - Allows a huge number of hyper-parameters with less over-fitting than MLE.
  - Can use gradient descent to optimize continuous hyper-parameters.
  - Ratio of marginal likelihoods (Bayes factor) can be used for hypothesis testing.
  - In many settings, naturally encourages sparsity (in parameters, data, clusters, etc.).

Beta-Bernoulli Model

- Consider again a coin-flipping example with a Bernoulli variable,

$$x \sim \text{Ber}(\theta).$$

- Previously we considered that either $\theta = 1$ or $\theta = 0.5$.

- Today: θ is a continuous variable coming from a beta distribution,

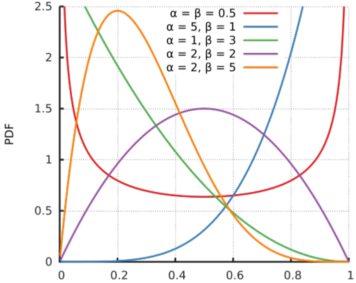
$$\theta \sim \mathcal{B}(\alpha, \beta).$$

- The parameters α and β of the prior are called hyper-parameters.
  - Similar to λ in regression, α and β are parameters of the prior.

Beta-Bernoulli Prior

Why the beta as a prior distribution?

- "It's a flexible distribution that includes uniform as special case".
- "It makes the integrals easy".



https://en.wikipedia.org/wiki/Beta_distribution

- Uniform distribution if $\alpha = 1$ and $\beta = 1$.
- "Laplace smoothing" corresponds to MAP with $\alpha = 2$ and $\beta = 2$.
  - Biased towards 0.5.

Beta-Bernoulli Posterior

- The PDF for the beta distribution has similar form to Bernoulli,

$$p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

- Observing HTH under Bernoulli likelihood and beta prior gives posterior of

$$\begin{aligned} p(\theta \mid HTH, \alpha, \beta) &\propto p(HTH \mid \theta, \alpha, \beta)p(\theta \mid \alpha, \beta) \\ &\propto \left(\theta^2(1-\theta)^1\theta^{\alpha-1}(1-\theta)^{\beta-1}\right) \\ &= \theta^{(2+\alpha)-1}(1-\theta)^{(1+\beta)-1}. \end{aligned}$$

- Since proportionality ($\propto$) constant is unique for probabilities, posterior is a beta:

$$\theta \mid HTH, \alpha, \beta \sim \mathcal{B}(2+\alpha, 1+\beta).$$

- When the prior and posterior come from same family, it's called a conjugate prior.

Conjugate Priors

- Conjugate priors make Bayesian inference easier:

1. Posterior involves updating parameters of prior.
    - For Bernoulli-beta, if we observe h heads and t tails then posterior is $\mathcal{B}(\alpha + h, \beta + t)$.
    - Hyper-parameters α and β are "pseudo-counts" in our mind before we flip.
2. We can update posterior sequentially as data comes in.
    - For Bernoulli-beta, just update counts h and t.

Conjugate Priors

- Conjugate priors make Bayesian inference easier:

3. Marginal likelihood has closed-form, proportional to ratio of normalizing constants.
   - The beta distribution is written in terms of the beta function B,

   $$p(\theta \mid \alpha, \beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \text{where} \quad B(\alpha,\beta) = \int_\theta \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta.$$

   and using the form of the posterior the marginal likelihood

   $$p(HTH \mid \alpha, \beta) = \int_\theta \frac{1}{B(\alpha,\beta)}\theta^{(h+\alpha)-1}(1-\theta)^{(t+\beta)-1}d\theta = \frac{B(h+\alpha, t+\beta)}{B(\alpha,\beta)}.$$

   - Empirical Bayes (type II MLE) would optimize this in terms of α and β.

4. In many cases posterior predictive also has a nice form...

Bernoulli-Beta Posterior Predictive

If we observe 'HHH' then our different estimates are:

- MAP with uniform Beta(1,1) prior (maximum likelihood),

$$\hat{\theta} = \frac{(3+\alpha)-1}{(3+\alpha)+\beta-2} = \frac{3}{3} = 1.$$

- MAP Beta(2,2) prior (Laplace smoothing),

$$\hat{\theta} = \frac{(3+\alpha)-1}{(3+\alpha)+\beta-2} = \frac{4}{6} = \frac{2}{3}.$$

Bernoulli-Beta Posterior Predictive

If we observe 'HHH' then our different estimates are:

- Posterior predictive (Bayesian) with uniform Beta(1,1) prior,

$$\begin{aligned} p(H \mid HHH) &= \int_0^1 p(H \mid \theta) p(\theta \mid HHH) d\theta \\ &= \int_0^1 \mathrm{Ber}(H \mid \theta) \mathrm{Beta}(\theta \mid 3+\alpha, \beta) d\theta \\ &= \int_0^1 \theta \mathrm{Beta}(\theta \mid 3+\alpha, \beta) d\theta = \mathbb{E}[\theta] \\ &= \frac{4}{5}. \end{aligned}$$

(mean of beta is $\alpha/(\alpha+\beta)$)

- Notice Laplace smoothing is not needed to avoid degeneracy under uniform prior.

Effect of Prior and Improper Priors

- We obtain different predictions under different priors:
  - $\mathcal{B}(3,3)$ prior is like seeing 3 heads and 3 tails (stronger prior towards 0.5),
    - For HHH, posterior predictive is 0.667.
  - $\mathcal{B}(100,1)$ prior is like seeing 100 heads and 1 tail (biased),
    - For HHH, posterior predictive is 0.990.
  - $\mathcal{B}(.01,.01)$ biases towards having unfair coin (head or tail),
    - For HHH, posterior predictive is 0.997.
    - Called "improper" prior (does not integrate to 1), but posterior can be "proper".
- We might hope to use an uninformative prior to not bias results.
  - But this is often hard/ambiguous/impossible to do (bonus slide).

Back to Conjugate Priors

- Basic idea of conjugate priors:

$$x \sim D(\theta), \quad \theta \sim P(\lambda) \quad \Rightarrow \quad \theta \mid x \sim P(\lambda').$$

- Beta-bernoulli example (beta is also conjugate for binomial and geometric):

$$x \sim \text{Ber}(\theta), \quad \theta \sim \mathcal{B}(\alpha, \beta), \quad \Rightarrow \quad \theta \mid x \sim \mathcal{B}(\alpha', \beta'),$$

- Gaussian-Gaussian example:

$$x \sim \mathcal{N}(\mu, \Sigma), \quad \mu \sim \mathcal{N}(\mu_0, \Sigma_0), \quad \Rightarrow \quad \mu \mid x \sim \mathcal{N}(\mu', \Sigma'),$$

  and posterior predictive is also a Gaussian.
- If Σ is also a random variable:
  - Conjugate prior is normal-inverse-Wishart, posterior predictive is a student t.
- For the conjugate priors of many standard distributions, see:

  https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions

Back to Conjugate Priors

- Conjugate priors make things easy because we have closed-form posterior.

- Some “non-named” conjugate priors:
  - Discrete priors are “conjugate” to all likelihoods:
    - Posterior will be discrete, although it still might be NP-hard to use.
  - Mixtures of conjugate priors are also conjugate priors.

- Do conjugate priors always exist?
  - No, they only exist for exponential family likelihoods (next slides).

- Bayesian inference is ugly when you leave exponential family (e.g., student t).
  - Can use numerical integration for low-dimensional integrals.
  - For high-dimensional integrals, need Monte Carlo methods or variational inference.

Digression: Exponential Family

- Exponential family distributions can be written in the form

$$p(x \mid w) \propto h(x) \exp(w^T F(x)).$$

- We often have $h(x) = 1$, or an indicator that x satisfies constraints.

- $F(x)$ is called the sufficient statistics.
  - $F(x)$ tells us everything that is relevant about data x.

- If $F(x) = x$, we say that the w are cannonical parameters.

- Exponential family distributions can be derived from maximum entropy principle.
  - Distribution that is "most random" that agrees with the sufficient statistics $F(x)$.
  - Argument is based on "convex conjugate" of $-\log p$.

Digression: Bernoulli Distribution as Exponential Family

- We often define linear models by setting $w^T x^i$ equal to cannonical parameters.
- If we start with the Gaussian (fixed variance), we obtain least squares.
- For Bernoulli, the cannonical parameterization is in terms of "log-odds",

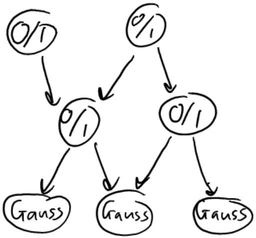
$$\begin{aligned} p(x \mid \theta) &= \theta^x(1-\theta)^{1-x} = \exp(\log(\theta^x(1-\theta)^{1-x})) \\ &= \exp(x\log\theta + (1-x)\log(1-\theta)) \\ &\propto \exp\left(x\log\left(\frac{\theta}{1-\theta}\right)\right). \end{aligned}$$

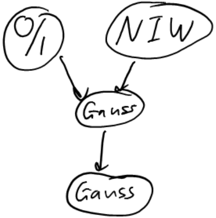
- Setting $w^T x^i = \log(y^i/(1-y^i))$ and solving for y^i yields logistic regression.
  - You can obtain regression models for other settings using this approach.

Conjugate Graphical Models

- DAG computations simplify if parents are conjugate to children.

- Examples:
  - Bernoulli child with Beta parent.
  - Gaussian belief networks.
  - Discrete DAG models.
  - Hybrid Gaussian/discrete, where discrete nodes can't have Gaussian parents.
  - Gaussian graphical model with normal-inverse-Wishart parents.





Summary

- Conjugate priors are priors that lead to posteriors of the same form.
  - They make Bayesian inference much easier.
- Exponential family distributions are the only distributions with conjugate priors.
- Next time: putting a prior on the prior and relaxing IID.

Uninformative Priors and Jeffreys Prior

- We might want to use an uninformative prior to not bias results.
  - But this is often hard/impossible to do.

- We might think the uniform distribution, $\mathcal{B}(1,1)$, is uninformative.
  - But posterior will be biased towards 0.5 compared to MLE.
  - And if you re-parameterize distribution it won't stay uniform.

- We might think to use "pseudo-count" of 0, $\mathcal{B}(0,0)$, as uninformative.
  - But posterior isn't a probability until we see at least one head and one tail.

- Some argue that the "correct" uninformative prior is $\mathcal{B}(0.5,0.5)$.
  - This prior is invariant to the parameterization, which is called a Jeffreys prior.