Probabilistic Modeling of Data (contd)

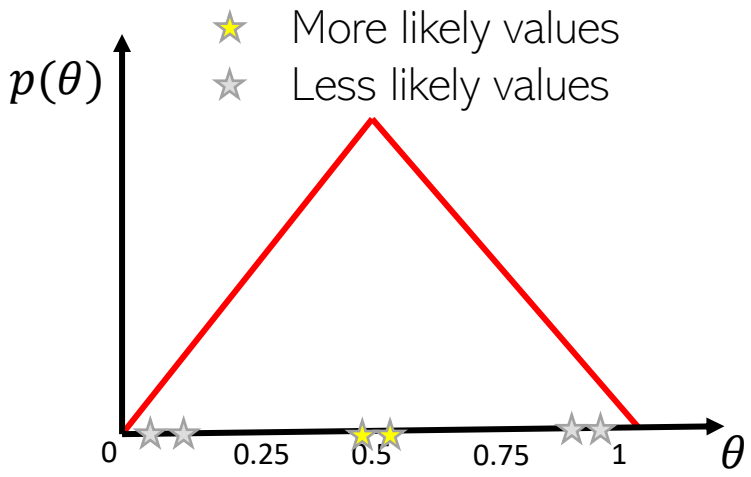
CS771: Introduction to Machine Learning

Plan today

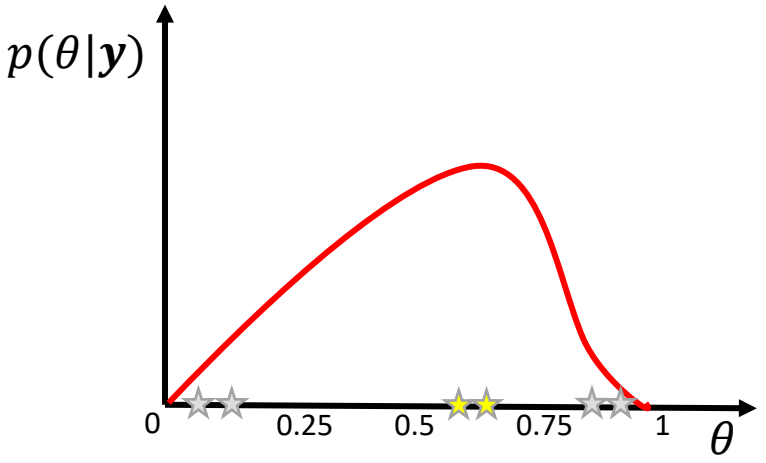
- Probabilistic modeling of data
  - MAP estimation
  - Bayesian inference (computing the posterior distribution)

Recap: Prior and Posterior

- Prior distribution reflects our prior belief about the unknown $\boldsymbol{\theta}$
- Posterior distribution reflects our updated belief once we have seen the data



$$p(\theta|\boldsymbol{y}) = \frac{p(\theta)p(\boldsymbol{y}|\theta)}{p(\boldsymbol{y})}$$



- Posterior's maxima is the MAP (maximum-a-posteriori) solution
- Posterior contains more information than just the MAP solution
  - It also tells us about the uncertainty in our estimate of $\boldsymbol{\theta}$

Recap: MLE and MAP

- MLE: Find the optimal parameter θ that maximizing the (log) likelihood

$$\theta_{MLE} = \operatorname*{argmax}_{\theta} \log p(\boldsymbol{y}|\theta) = \operatorname*{argmax}_{\theta} \sum_{n=1}^{N} \log p(y_n|\theta)$$

- MAP: Find the optimal parameter θ that has the maximum posterior probability

$$\theta_{MAP} = \operatorname*{argmax}_{\theta} \log p(\theta|\boldsymbol{y}) = \operatorname*{argmax}_{\theta} [\log p(\boldsymbol{y}|\theta) + \log p(\theta)]$$

$$= \operatorname*{argmax}_{\theta} \left[\sum_{n=1}^{N} \log p(y_n|\theta) + \log p(\theta)\right]$$

Important: Computing the MAP solution does not require computing the posterior (this maximization can be done even without computing the posterior)!

- MAP's advantage over MLE: The prior acts as a regularizer on θ

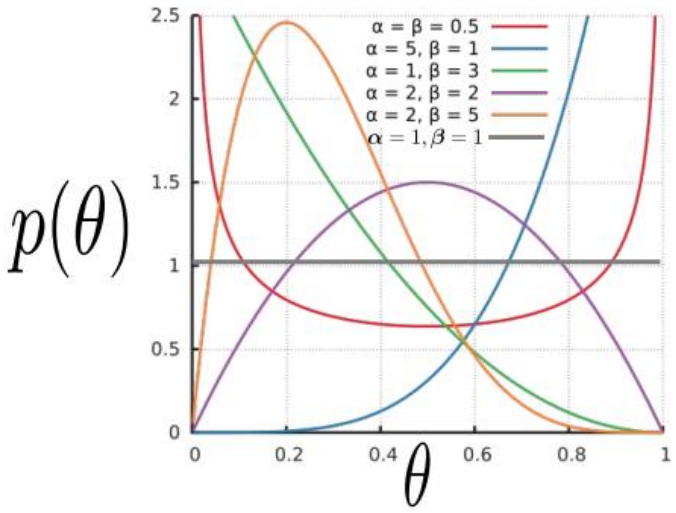
- Negative log likelihood is akin to loss function, negative log prior is akin to regularizer

MAP Estimation: An Example

- Let's again consider the coin-toss problem (estimating the bias of the coin)
- Each likelihood term is Bernoulli

$$p(y_n|\theta) = \text{Bernoulli}(y_n|\theta) = \theta^{y_n}(1-\theta)^{1-y_n}$$

- Also need a prior since we want to do MAP estimation
- Since $\theta \in (0,1)$, a reasonable choice of prior for θ would be Beta distribution



$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

The gamma function

Using $\alpha = 1$ and $\beta = 1$ will make the Beta prior a uniform prior

α and β (both non-negative reals) are the two hyperparameters of this Beta prior

Can set these based on intuition, cross-validation, or even learn them

MAP Estimation: An Example (Contd)

- The log posterior for the coin-toss model is log-lik + log-prior

$$LP(\theta) = \sum_{n=1}^{N} \log p(y_n|\theta) + \log p(\theta|\alpha, \beta)$$

- Plugging in the expressions for Bernoulli and Beta and ignoring any terms that don't depend on θ, the log posterior simplifies to

$$LP(\theta) = \sum_{n=1}^{N} [y_n \log \theta + (1 - y_n)\log(1 - \theta)] + (\alpha - 1)\log \theta + (\beta - 1)\log(1 - \theta)$$

- Maximizing the above log post. (or min. of its negative) w.r.t. θ gives

$$\theta_{MAP} = \frac{\sum_{n=1}^{N} y_n + \alpha - 1}{N + \alpha + \beta - 2}$$

Using $\alpha = 1$ and $\beta = 1$ gives us the same solution as MLE

Recall that $\alpha = 1$ and $\beta = 1$ for Beta distribution is in fact equivalent to a uniform prior (hence making MAP equivalent to MLE)

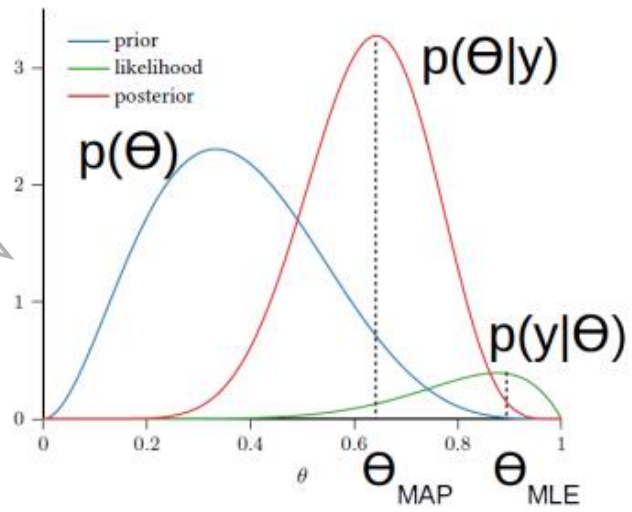
Prior's hyperparameters have an interesting interpretation. Can think of $\alpha - 1$ and $\beta - 1$ as the number of heads and tails, respectively, before starting the coin-toss experiment (akin to "pseudo-observations")

Such interpretations of prior's hyperparameters as being "pseudo-observations" exist for various other prior distributions as well (in particular, distributions belonging to "exponential family" of distributions

Fully Bayesian Inference

- MLE/MAP only give us a point estimate of θ

MAP estimate is more robust than MLE (due to the regularization effect) but the estimate of uncertainty is missing in both approaches – both just return a single "optimal" solution by solving an optimization problem



Interesting fact to keep in mind: Note that the use of the prior is making the MLE solution move towards the prior (MAP solution is kind of a "compromise between MLE solution of the mode of the prior) ☺

Fully Bayesian inference

- If we want more than just a point estimate, we can compute the full posterior

$$p(\theta|\boldsymbol{y}) = \frac{p(\theta)p(\boldsymbol{y}|\theta)}{p(\boldsymbol{y})}$$

Computable analytically only when the prior and likelihood are "friends" with each other (i.e., they form a conjugate pair of distributions (distributions from exponential family have conjugate priors

An example: Bernoulli and Beta are conjugate. Will see some more such pairs

In other cases, the posterior needs to be approximated (will see 1-2 such cases in this course; more detailed treatment in the advanced course on probabilistic modeling and inference)

Fully Bayesian Inference: An Example

- Let's again consider the coin-toss problem
- Bernoulli likelihood: $p(y_n|\theta) = \text{Bernoulli}(y_n|\theta) = \theta^{y_n}(1-\theta)^{1-y_n}$
- Beta prior: $p(\theta) = \text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- The posterior can be computed as

Also, if you get more observations, you can treat the current posterior as the new prior and obtain a new posterior using these extra observations

Posterior in this example is the same distribution as the prior (both Beta), just with updated hyperparameters (property when likelihood and prior are conjugate to each other)

Number of heads (N_1)

Number of tails (N_0)

$$\theta^{\sum_{n=1}^{N} y_n}(1-\theta)^{N-\sum_{n=1}^{N} y_n}$$

$$p(\theta|\boldsymbol{y}) = \frac{p(\theta)p(\boldsymbol{y}|\theta)}{p(\boldsymbol{y})} = \frac{p(\theta)\prod_{n=1}^{N} p(y_n|\theta)}{p(\boldsymbol{y})} = \frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{n=1}^{N}\theta^{y_n}(1-\theta)^{1-y_n}}{\int \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{n=1}^{N}\theta^{y_n}(1-\theta)^{1-y_n} d\theta}$$

This is the numerator integrated/marginalized over θ : $p(\mathbf{y}) = \int p(\theta,\boldsymbol{y})d\theta = \int p(\theta)p(\boldsymbol{y}|\theta)d\theta$

In general, hard but with conjugate pairs of prior and likelihood, we don't need to compute this, as we will see in this example ☺

Parts coming from the numerator, which consist of θ terms. We have ignored other constants in the numerator, and the whole denominator which is also constant w.r.t. θ

$$\propto \theta^{\alpha+N_1-1}(1-\theta)^{\beta+N_0-1}$$

Aha! This is nothing but

$$\text{Beta}(\theta|\alpha+N_1,\beta+N_0)$$

This, of course, is not always possible but only in simple cases like this

Found the posterior just by simple inspection without having to calculate the constant of proportionality ☺

Conjugacy

- Many pairs of distributions are conjugate to each other
  - Bernoulli (likelihood) + Beta (prior) ⇒ Beta posterior
  - Binomial (likelihood) + Beta (prior) ⇒ Beta posterior
  - Multinomial (likelihood) + Dirichlet (prior) ⇒ Dirichlet posterior
  - Poisson (likelihood) + Gamma (prior) ⇒ Gamma posterior
  - Gaussian (likelihood) + Gaussian (prior) ⇒ Gaussian posterior
  - and many other such pairs ..
- Tip: If two distr are conjugate to each other, their functional forms are similar
  - Example: Bernoulli and Beta have the forms

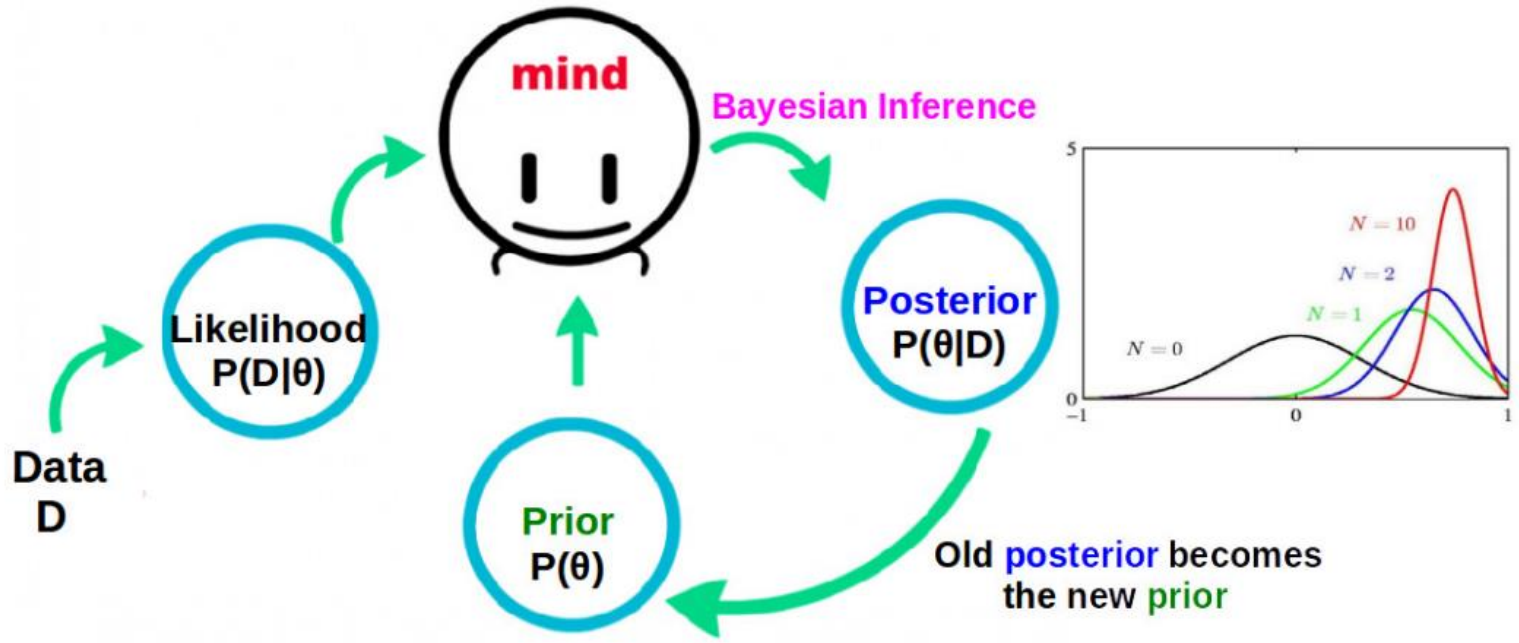
$$\text{Bernoulli}(y|\theta) = \theta^{y}\,(1-\theta)^{1-y}$$

$$\text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

This is why, when we multiply them while computing the posterior, the exponents get added and we get the same form for the posterior as the prior but with just updated hyperparameter. Also, we can identify the posterior and its hyperparameters simply by inspection

"Online" Nature of Bayesian Inference

- Fully Bayesian inference fits naturally into an "online" learning setting



Also, the posterior becomes more and more "concentrated" as the number of observations increases. For very large N, you may expect it to be peak around the MLE solution

- Our belief about θ keeps getting updated as we see more and more data

Probabilistic Models: Making Predictions

- Having estimated $\boldsymbol{\theta}$ (MLE/MAP/posterior), we can now use it to make predictions
- Prediction entails computing posterior predictive distribution of a new observation $\boldsymbol{y}_*$

$$p(y_*|\boldsymbol{y}) = \int p(y_*, \theta|\boldsymbol{y}) d\theta$$

Marginalizing (summing/integrating) over the unknown random variable $\boldsymbol{\theta}$

$$= \int p(y_*|\theta, \boldsymbol{y}) p(\theta|\boldsymbol{y}) d\theta$$

Decomposing the joint using chain rule

Conditional distribution of the new observation, given past observations

$$= \int p(y_*|\theta) p(\theta|\boldsymbol{y}) d\theta$$

This step assumes i.i.d. data, i.e., given $\boldsymbol{\theta}$, $\boldsymbol{y}_*$ does not depend on $\boldsymbol{y}$

In some simple cases, this can be computed exactly but, in general, it can't be and needs to be approximated

$$= \mathbb{E}_{p(\theta|\boldsymbol{y})}[p(y_*|\theta)]$$

This computes the predictive distribution by averaging over the full posterior – basically calculates $p(y_*|\theta)$ for each possible θ, weighs it by the probability of θ under the posterior $p(\theta|\boldsymbol{y})$, and sums all such posterior weighted predictions. Note that not each value of θ is given equal importance here in the averaging

- When doing MLE/MAP, we approximate the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ by a single point $\boldsymbol{\theta}_{opt}$

$$p(y_*|\boldsymbol{y}) = \int p(y_*|\theta) p(\theta|\boldsymbol{y}) d\theta \approx p(y_*|\theta_{opt})$$

A "plug-in predictive distribution" - simply plugged in the single best estimate (MLE/MAP) that we have

- Plug-in prediction which uses MLE/MAP of $\boldsymbol{\theta}$ is cheaper since no integral involved

Making Predictions: An Example

- For coin-toss example, prediction means computing $p(y_{N+1} = 1|\boldsymbol{y})$
- This can be done using the MLE/MAP estimate, or using the full posterior

$$\theta_{MLE} = \frac{N_1}{N} \qquad \theta_{MAP} = \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2} \qquad p(\theta|\boldsymbol{y}) = \text{Beta}(\theta|\alpha + N_1, \beta + N_0)$$

- Thus for this example (where observations are assumed iid from a Bernoulli)

$$\text{MLE prediction:} \quad p(y_{N+1} = 1|\boldsymbol{y}) = \int p(y_{N+1} = 1|\theta)p(\theta|\boldsymbol{y})d\theta \approx p(y_{N+1} = 1|\theta_{MLE}) = \theta_{MLE} = \frac{N_1}{N}$$

$$\text{MAP prediction:} \quad p(y_{N+1} = 1|\boldsymbol{y}) = \int p(y_{N+1} = 1|\theta)p(\theta|\boldsymbol{y})d\theta \approx p(y_{N+1} = 1|\theta_{MAP}) = \theta_{MAP} = \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2}$$

$$\text{Fully Bayesian:} \quad p(y_{N+1} = 1|\boldsymbol{y}) = \int p(y_{N+1} = 1|\theta)p(\theta|\boldsymbol{y})d\theta = \int \theta p(\theta|\boldsymbol{y})d\theta = \int \theta \text{Beta}(\theta|\alpha + N_1, \beta + N_0)d\theta = \frac{N_1 + \alpha}{N + \alpha + \beta}$$

Again, keep in mind that the posterior weighted averaged prediction used in the fully Bayesian case would usually not be as simple to compute as it was in this case. We will look at some hard cases later

Expectation of θ under the Beta posterior that we computed using fully Bayesian inference

Probabilistic Modeling: A Summary

- Likelihood corresponds to a loss function; prior corresponds to a regularizer
- Can choose likelihoods and priors based on the nature/property of data/parameters
- MLE estimation = unregularized loss function minimization
- MAP estimation = regularized loss function minimization
- Allows us to do fully Bayesian learning (learning the full distribution of the parameters)
- Makes robust predictions by posterior averaging (rather than using point estimate)
- Many other benefits, such as
  - Estimate of confidence in the model's prediction (useful for doing Active Learning)
  - Can do automatic model selection, hyperparameter estimation, handle missing data, etc.
  - Formulate latent variable models
  - .. and many other benefits (a proper treatment deserves a separate course, but we will see some of these in this course, too)