COMPSCI 689
Lecture 20: Introduction to Bayesian Learning

Brendan O'Connor

College of Information and Computer Sciences
University of Massachusetts Amherst

Slides by Benjamin M. Marlin (marlin@cs.umass.edu).

Learning and Probabilistic Models

- In the previous two units of the course, we focused on an optimization approach to learning.
- In both the MLE and ERM frameworks, we have tried to identify the optimal parameters θ of a probability distribution or prediction function given a sample of data $\mathcal{D}$.
- However, when the amount of data is low relative to the number of parameters, such approaches tend to overfit the available data and perform poorly.
- The optimization approach to dealing with this problem is to apply regularization or penalization.

Parameter Uncertainty

- The underlying problem with MLE and ERM when using small data sets $\mathcal{D}$ is that the resulting approximation $P_{\mathcal{D}}$ of the true data generating distribution P_* can be very poor.
- An alternative view on this situation is that based on a small data set $\mathcal{D}$, there can be substantial uncertainty about the correct value of the unknown parameters θ.
- This view motivates the idea of treating the unknown parameters θ as an additional set of random variables.
- This idea is the foundational concept of Bayesian statistics. Instead of trying to find the best single choice of the parameters θ given the data, we treat the parameters θ as an extra set of latent variables and marginalize over them.

Definitions I

- Parameters: $\theta \in \mathbb{R}^K$
- Data (Unsupervised): $\mathcal{D} = \{\mathbf{x}_n\}_{1:N}$
- Likelihood (Unsupervised): $P(\mathcal{D}|\theta) = \prod_{n=1}^{N} P(\mathbf{X}_n = \mathbf{x}_n|\theta)$
- Data (Supervised): $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{1:N}$
- Conditional Likelihood (Supervised): $P(\mathcal{D}|\theta) = \prod_{n=1}^{N} P(Y_n = y_n|\mathbf{X}_n = \mathbf{x}_n, \theta)$

Definitions II

- Prior: $P(\theta)$
- Joint: $P(\mathcal{D}, \theta) = P(\mathcal{D}|\theta)P(\theta)$
- Evidence: $P(\mathcal{D}) = \int P(\mathcal{D}, \theta)d\theta = \int P(\mathcal{D}|\theta)P(\theta)d\theta$
- Parameter Posterior: $P(\theta|\mathcal{D}) = \frac{P(\mathcal{D},\theta)}{P(\mathcal{D})} = \frac{P(\mathcal{D}|\theta)P(\theta)}{\int P(\mathcal{D}|\theta)P(\theta)d\theta}$
- Posterior Predictive Distribution (Unsupervised): $P(\mathbf{x}|\mathcal{D}) = \int P(\mathbf{x}|\theta)P(\theta|\mathcal{D})d\theta$
- Posterior Predictive Distribution (Supervised): $P(y|\mathbf{x}, \mathcal{D}) = \int P(y|\mathbf{x}, \theta)P(\theta|\mathcal{D})d\theta$

Bayesian Inference

- The term *Bayesian inference* refers specifically to the problem of computing the parameter posterior $P(\theta|\mathcal{D})$, which can be thought of as a process of updating the prior $P(\theta)$ to account for the available data $\mathcal{D}$.
- Instead of identifying the single best parameter according to some criterion, we maintain a probability distribution over the unknown value of θ.
- Note that this process supports incremental updating. We convert our current prior into a parameter posterior given our current data. If we obtain more data, we can use our current parameter posterior as a prior, incorporate the new data, and obtain a new parameter posterior.

Bayesian Prediction

- Prediction in Bayesian statistics involves the complete parameter posterior:
  $P(\mathbf{x}|\mathcal{D}) = \int P(\mathbf{X} = \mathbf{x}|\theta)P(\theta|\mathcal{D})d\theta$
  $P(y|\mathbf{x}, \mathcal{D}) = \int P(y|\mathbf{x}, \theta)P(\theta|\mathcal{D})d\theta.$
- As we can see, the parameters are unknown and thus must be marginalized away. The only quantity we condition on is the observed data $\mathcal{D}$.
- The distributions $P(\mathbf{x}|\mathcal{D})$ and $P(y|\mathbf{x}, \mathcal{D})$ can be viewed as continuous mixtures, averaging over all possible model parameters weighted by their posterior probabilities.
- This is called *Bayesian model averaging* and is closely related to ensemble methods.

Plug-In Estimates

- When the distribution $P(\theta|\mathcal{D})$ concentrates its mass on a small region of parameter space, summarizing the full posterior by a delta function centered at the mode of the posterior can yield a reasonable approximation:

$$\theta_* = \arg\max_\theta P(\theta|\mathcal{D}) \;=\; \arg\max_\theta \log P(\mathcal{D}|\theta) + \log P(\theta)$$

- This is referred to as *maximum a posteriori* (MAP) estimation. The learning problem is equivalent to regularized MLE.

- Plugging this approximation into the posterior predictive distribution is called a *plug-in* estimate and yields:

$$\begin{aligned} P(\mathbf{x}|\mathcal{D}) &= \int P(\mathbf{X}=\mathbf{x}|\theta)P(\theta|\mathcal{D})d\theta \\ &\approx \int P(\mathbf{X}=\mathbf{x}|\theta)\delta(\theta,\theta_*)d\theta = P(\mathbf{X}=\mathbf{x}|\theta_*) \end{aligned}$$

Plug-In Estimates

- This derivation makes it very clear when use of the MAP approximation is well justified from a Bayesian perspective.
- If $P(\theta|\mathcal{D})$ is not sufficiently concentrated around θ_*, then we are strictly discarding uncertainty in the parameters when we make predictions using $P(\mathbf{X} = \mathbf{x}|\theta_*)$ and we should expect lower predictive likelihoods.
- The same is true in the case of the MLE. If we compute the MLE as: $\theta_* = \arg\max_\theta \log P(\mathcal{D}|\theta)$, and then use it to make predictions, we need enough data that $P(\theta|\mathcal{D})$ concentrates around θ_* or we are again discarding uncertainty.

Plug-In Estimates

- The concentration of $P(\theta|\mathcal{D})$ around the MLE will eventually occur (under regularity conditions), but it requires sufficient data for the effect of the prior to go to zero.
- When sufficient data are not available, the MLE and MAP can differ substantially from each other, and both can be very poor summaries of $P(\theta|\mathcal{D})$.

When not to use Bayesian Inference

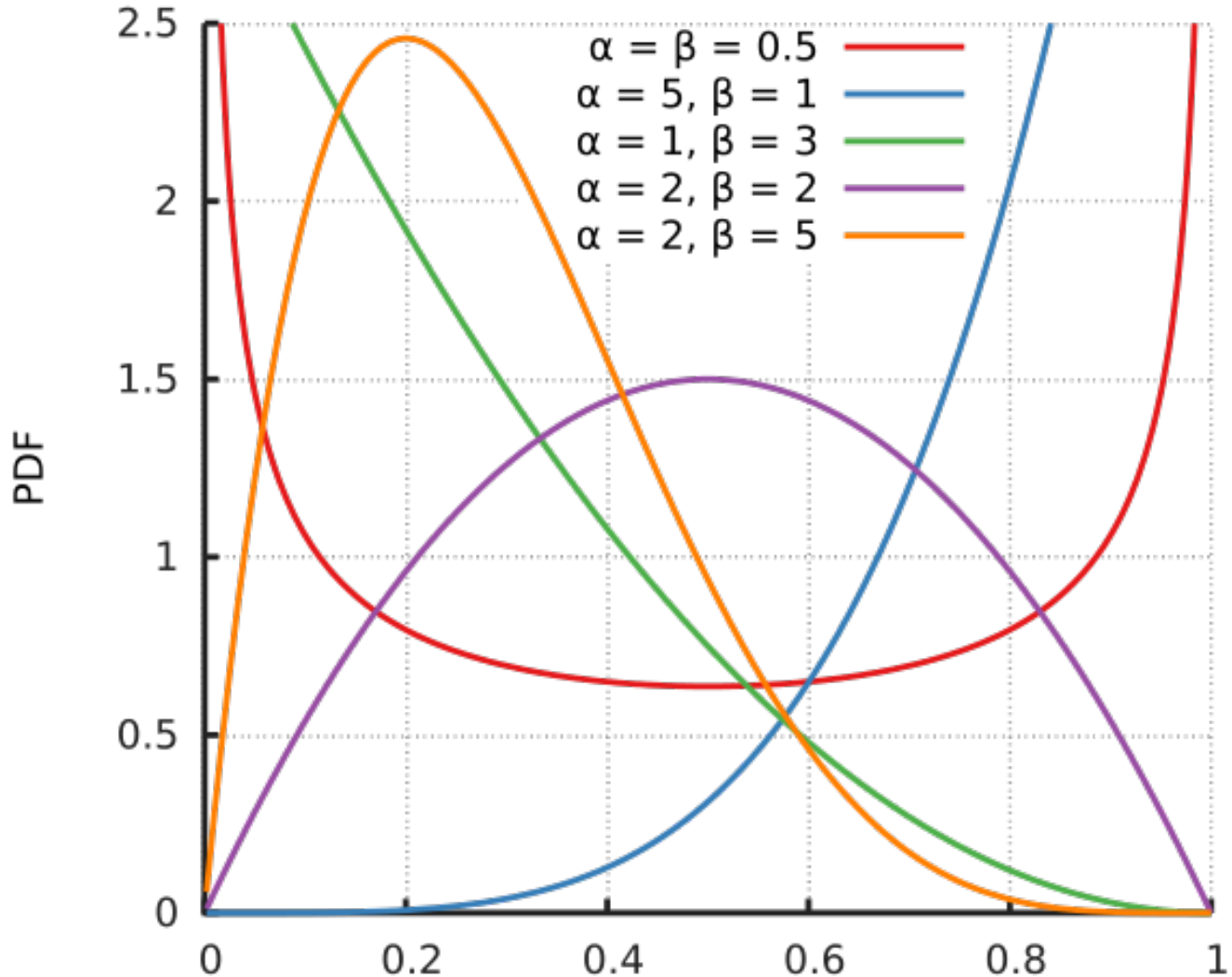
- The flip side of the previous argument is that when N is large, there is no point in doing the extra work required by Bayesian inference because the parameter posterior will converge to a delta function at the MAP estimate of θ.

The Beta-Bernoulli Model

- The Beta-Bernoulli model is a classical application of the ideas of Bayesian inference.
- This model has binary data $x \in \{0, 1\}$ with a Bernoulli likelihood $P(\mathcal{D}|\theta) = \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{(1-x_n)}$.
- The prior distribution on the unknown parameter $\theta \in [0, 1]$ is given by the Beta distribution:

$$P(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

The Beta Distribution



The Beta Distribution

- The Beta distribution is the *conjugate prior* to the Bernoulli likelihood. A conjugate prior has the same functional form as the terms in the likelihood that involve the parameters.
- The use of a conjugate prior results in a posterior that belongs to the same family of distributions as the prior.
- The function $\Gamma(x)$ in the definition of the Beta distribution is called the *gamma function* and is a generalization of the factorial function to the real numbers. For positive reals, it satisfies the property $\Gamma(x+1) = x\Gamma(x)$.
- Note that since we know that $P(\theta)$ is a normalized probability density, we have that $\int P(\theta)d\theta = 1$, and thus:

$$\int \theta^{a-1}(1-\theta)^{b-1}d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Beta-Bernoulli Inference: Joint

Likelihood: simplify using N_1, N_0: number of 1s and 0s.

$$P(\mathcal{D}|\theta) = \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{(1-x_n)} = \theta^{N_1}(1-\theta)^{N_0}$$

The joint distribution is then given by:

$$\begin{aligned} P(\mathcal{D},\theta) &= \theta^{N_1}(1-\theta)^{N_0} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \end{aligned}$$

Beta-Bernoulli Inference: Evidence

To get the evidence, we marginalize the joint over the parameters. This uses the unnormalized Beta integral result we noted above:

$$P(\mathcal{D}) = \int P(\mathcal{D}, \theta) d\theta$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} d\theta$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(N_1+a)\Gamma(N_0+b)}{\Gamma(N+a+b)}$$

Beta-Bernoulli Inference: Parameter Posterior

Now, we can get the parameter posterior as shown below:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D},\theta)}{P(\mathcal{D})}$$

$$= \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{N_1+a-1}(1-\theta)^{N_0+b-1}}{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(N_1+a)\Gamma(N_0+b)}{\Gamma(N+a+b)}}$$

$$= \frac{\Gamma(N+a+b)}{\Gamma(N_1+a)\Gamma(N_0+b)}\theta^{N_1+a-1}(1-\theta)^{N_0+b-1}$$

$$= \text{Beta}(N_1+a, N_0+b)$$

Beta-Bernoulli Inference: Posterior Predictive

Finally, we can obtain the posterior predictive distribution:

$$
\begin{aligned}
P(X=1|\mathcal{D}) &= \int P(X=1|\theta)\cdot P(\theta|\mathcal{D})d\theta \\
&= \int \theta \cdot P(\theta|\mathcal{D})d\theta \\
&= \int \theta \cdot \frac{\Gamma(N+a+b)}{\Gamma(N_1+a)\Gamma(N_0+b)}\theta^{N_1+a-1}(1-\theta)^{N_0+b-1}d\theta \\
&= \frac{\Gamma(N+a+b)}{\Gamma(N_1+a)\Gamma(N_0+b)}\int \theta^{N_1+a+1-1}(1-\theta)^{N_0+b-1}d\theta \\
&= \frac{\Gamma(N+a+b)}{\Gamma(N_1+a)\Gamma(N_0+b)}\frac{\Gamma(N_1+a+1)\Gamma(N_0+b)}{\Gamma(N+a+b+1)}
\end{aligned}
$$

Beta-Bernoulli Inference: Posterior Predictive

$$
\begin{aligned}
P(X=1|\mathcal{D}) &= \frac{\Gamma(N+a+b)}{\Gamma(N_1+a)\Gamma(N_0+b)} \frac{\Gamma(N_1+a+1)\Gamma(N_0+b)}{\Gamma(N+a+b+1)} \\
&= \frac{\Gamma(N+a+b)}{\Gamma(N_1+a)\Gamma(N_0+b)} \frac{\Gamma(N_1+a)(N_1+a)\Gamma(N_0+b)}{\Gamma(N+a+b)(N+a+b)} \\
&= \frac{N_1+a}{N+a+b}
\end{aligned}
$$

Analysis

- Given N_1 ones and N_0 zeros in the data set, the posterior predictive distribution gives the probability that the next observation will be a one as $\frac{N_1+a}{N+a+b}$.
- If we use the MLE as a plug-in estimate, we obtain $\theta_{MLE} = N_1/N$.
- If we compute the posterior mode, we would find $\theta_{MAP} = \frac{N_1+a-1}{N+a+b-2}$.

Analysis

- Further, we can see that as N goes to infinity, the effect of the prior will go to zero as expected.
- However, when N is small, the three predictive probabilities can differ substantially.
- In general, the MLE will be the worst conditioned. With only one observation in the data set, the MLE predicts that all future data cases will match the first observation with probability 1. This is nonsense (almost always).
- The MAP will be better conditioned, but only if a and b are strictly greater than one.
- The Bayesian posterior predictive distribution makes a non-degenerate prediction when a and b are any non-zero values.

Bayesian Linear Regression

Remember the probabilistic model interpretation of ridge regression:

$$p(w) \sim N(0, \lambda^{-1} I_D)$$
$$p(y \mid X, w, \mu, \sigma^2) \sim N(\mu + Xw, \sigma^2 I_N)$$

(drawing entire dataset y from one N-dimensional MVN)
L2-regularized lin.reg. is MAP estimation for $p(w \mid X, y, \lambda, \sigma^2)$. But it turns out the full w posterior isa closed form MVN as well! Due to linear Gaussian conjugacy, deriving from conditioning rules of MVNs (last week).

Posterior

$$p(w \mid X, y, \sigma^2) \propto N(w \mid 0, \lambda^{-1})N(y \mid Xw, \sigma^2 I) = N(w \mid w_N, V_N)$$

$$w_N = \frac{1}{\sigma^2} V_N X^T y$$

$$V_N^{-1} = \lambda + \frac{1}{\sigma^2} X^T X$$

Example

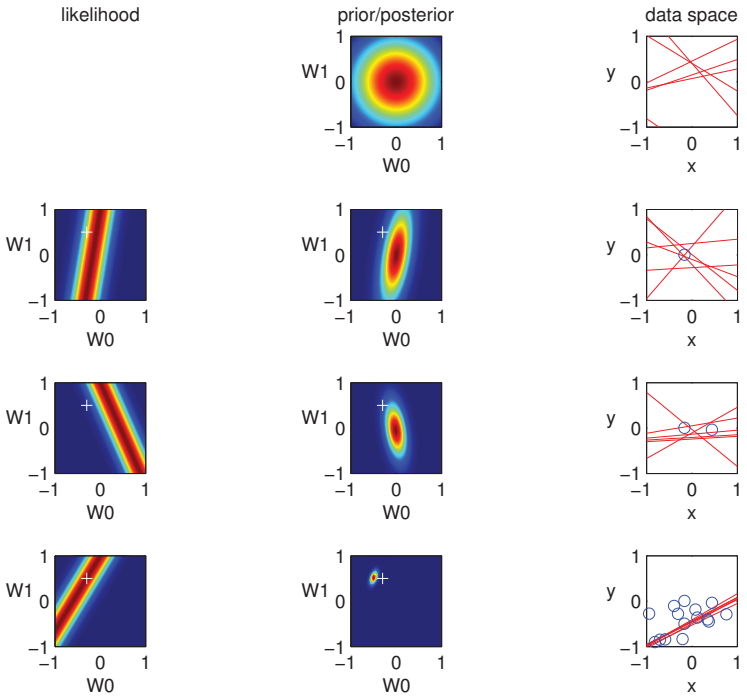
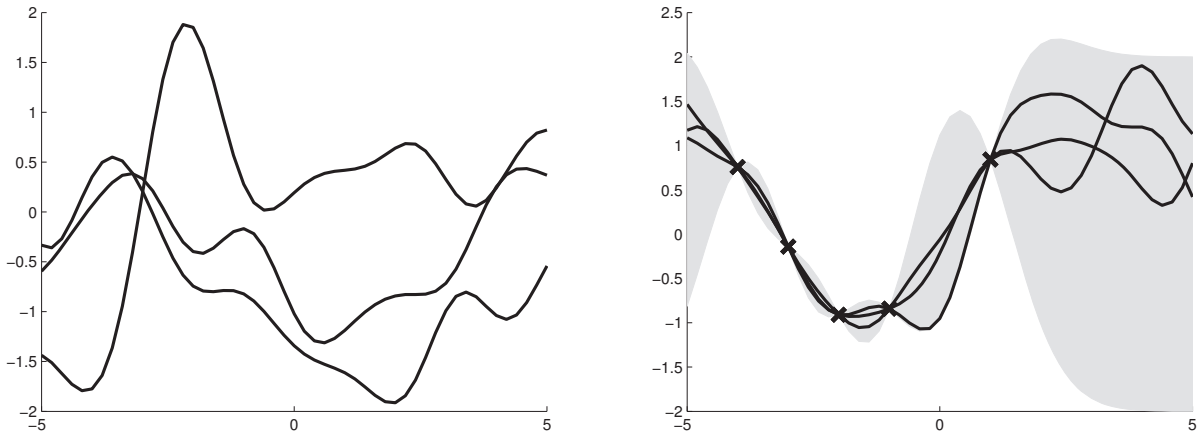


Figure 7.11 Sequential Bayesian updating of a linear regression model $p(y|\mathbf{x}) = \mathcal{N}(y|w_0x_0 + w_1x_1, \sigma^2)$. Row 0 represents the prior, row 1 represents the first data point (x_1, y_1), row 2 represents the second data point (x_2, y_2), row 3 represents the 20th data point (x_{20}, y_{20}). Left column: likelihood function for current data point. Middle column: posterior given data so far, $p(\mathbf{w}|\mathbf{x}_{1:n}, y_{1:n})$ (so the first line is the prior). Right column: samples from the current prior/posterior predictive distribution. The white cross in columns 1 and 2 represents the true parameter value; we see that the mode of the posterior rapidly (after 20 samples) converges to this point. The blue circles in column 3 are the observed data points. Based on Figure 3.7 of (Bishop 2006a). Figure generated by `bayesLinRegDemo2d`.

Gaussian Processes

We won't cover this in detail. But you can kernelize Bayesian linear regression, with closed-form sampling of prediction functions and excellent uncertainty estimation. Downside: quadratic to cubic in training set size and/or number features (similar to other kernel methods).
But usually models don't have closed-form Bayesian posteriors. Next time, approximate inference.

Example



(a) (b)

Figure 15.2 Left: some functions sampled from a GP prior with SE kernel. Right: some samples from a GP posterior, after conditioning on 5 noise-free observations. The shaded area represents $\mathbb{E}[f(\mathbf{x})] \pm 2\text{std}(f(\mathbf{x}))$. Based on Figure 2.2 of (Rasmussen and Williams 2006). Figure generated by `gprDemoNoiseFree`.