Bayesian Learning
SC4/SM4 Data Mining and Machine Learning, Hilary Term 2017
Dino Sejdinovic

7.1 Bayesian Inference

So far, our treatment of probabilistic machine learning models has been *frequentist*, i.e. we used one set of tools to reason about latent variables $\mathbf{z}$ (e.g. cluster indicators in a mixture model) and another to reason about model parameters θ (e.g. parameters of mixture components defining those clusters). The generative processes we considered define the *likelihood function*: the joint distribution $p(\mathcal{D}|\theta)$ of all the observed data $\mathcal{D}$ given the model parameters θ and the learning consists in computing the maximum likelihood estimator

$$\hat{\theta} = \arg\max_{\theta \in \Theta} p(\mathcal{D}|\theta).$$

For example, in the EM algorithm (which is a frequentist method aimed at locally maximising the likelihood function), we were placing a variational distribution q on latent variables but not on θ, which was inferred using point estimates at each iteration.

In Bayesian inference, we also treat the model parameters θ as random variables and the process of learning is then computation of the *posterior distribution* $p(\theta|\mathcal{D})$. In addition to the likelihood $p(\mathcal{D}|\theta)$ specified by the generative model, one needs to also specify a prior distribution $p(\theta)$. Posterior distribution is then given by the *Bayes Theorem*:

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

where the denominator is the *marginal likelihood* or *evidence*:

$$p(\mathcal{D}) = \int_{\Theta} p(\mathcal{D}|\theta)p(\theta)d\theta.$$

All the questions about model parameters can be addressed based on the posterior. We can, for example, consider

- *Posterior mode*: $\widehat{\theta}^{\mathrm{MAP}} = \arg\max_{\theta\in\Theta} p(\theta|\mathcal{D})$ (maximum a posteriori).
- *Posterior mean*: $\widehat{\theta}^{\mathrm{mean}} = \mathbb{E}[\theta|\mathcal{D}]$.
- *Posterior variance*: $\mathrm{Var}[\theta|\mathcal{D}]$.
- *Posterior expectations of functions of parameters*: $\mathbb{E}[g(\theta)|\mathcal{D}]$ for some $g : \Theta \to \mathbb{R}^s$.

A particularly convenient choice of prior distributions are *conjugate priors* to a given likelihood function. A prior and likelihood are said to be conjugate if they result in a posterior that lies in the same parametric family as the prior.

Example: Bayesian inference on a categorical distribution. Suppose we observe $\mathcal{D} = \{y_i\}_{i=1}^n$, with $y_i \in \{1, \ldots, K\}$, and model them as i.i.d. with the probability mass function $\pi = (\pi_1, \ldots, \pi_K)$:

$$p(\mathcal{D}|\pi) = \prod_{i=1}^{n} \pi_{y_i} = \prod_{k=1}^{K} \pi_k^{n_k}$$

with $n_k = \sum_{i=1}^n \mathbf{1}(y_i = k)$ and $\pi_k > 0$, $\sum_{k=1}^K \pi_k = 1$. The conjugate prior on π is the Dirichlet distribution $\mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$ with parameters $\alpha_k > 0$, and density

$$p(\pi) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

on the probability simplex $\{\pi : \pi_k > 0, \sum_{k=1}^K \pi_k = 1\}$. Since

$$p\left(\pi|\mathcal{D}\right) \quad \propto \quad \prod_{k=1}^{K} \pi_k^{n_k + \alpha_k - 1},$$

the posterior is also Dirichlet $\mathrm{Dir}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$. Posterior mean is given by

$$\widehat{\pi}_k^{\text{mean}} = \frac{\alpha_k + n_k}{\sum_{j=1}^K \alpha_j + n_j}.$$

Notice how parameters of the prior (hyperparameters) are essentially playing the role of the pseudocounts for each of the classes $1, \ldots, K$ (but they need not be integer-valued). They are reflecting prior beliefs about class proportions. For the case of two classes, this is equivalent to a Beta (α_1, α_2) prior on π_1, i.e. $p(\pi_1) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_1)} \pi_1^{\alpha_1} (1 - \pi_1)^{\alpha_2}$.

7.2 Predictive distributions

How do we construct predictions based on the posterior distributions? Write the observations as $\mathcal{D} = \{x_i\}_{i=1}^n$ and assume the generative model specifies $p(x|\theta)$, e.g. a mixture model $p(x|\theta) = \sum_{k=1}^K \pi_k f\left(x|\phi_k\right)$, with $\theta = (\pi_1, \ldots, \pi_K; \phi_1, \ldots, \phi_K)$. The *posterior predictive distribution* is the conditional distribution of x_{n+1} given $\mathcal{D} = \{x_i\}_{i=1}^n$:

$$\begin{aligned} p(x_{n+1}|\mathcal{D}) &= \int_\Theta p(x_{n+1}|\theta, \mathcal{D}) p(\theta|\mathcal{D}) d\theta \\ &= \int_\Theta p(x_{n+1}|\theta) p(\theta|\mathcal{D}) d\theta. \end{aligned}$$

Thus, we predict new data by *averaging the predictive distribution over the posterior*. This is fundamentally different than predicting using a point estimate of θ, i.e. $p(x_{n+1}|\hat{\theta})$ as it takes into account the posterior uncertainty in parameters.

Example: Bayesian treatment of naïve Bayes classifer. Consider a K-class classification problem with binary input vectors, i.e. $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, $x_i \in \{0,1\}^p$ and $y_i \in \{1, \ldots, K\}$. Naïve Bayes classifier uses the following model:

$$p(y_i = k|\theta) = \pi_k, \quad p(x_i|y_i = k, \theta) = \prod_{j=1}^{p} \phi_{kj}^{x_i^{(j)}} (1 - \phi_{kj})^{1 - x_i^{(j)}},$$

i.e. it assumes that given the class labels, individual dimensions in input vectors are independent. The parameters of the model are collated into $\theta = ((\pi_k), (\phi_{kj}))$. It is often used in text classification where data items correspond to documents and $x_i^{(j)}$ indicates whether word j from a list of p words has appeared in document i. Class labels correspond to e.g. topics of the documents. Despite the name, naïve Bayes is often treated in a frequentist way, i.e. using maximum likelihood estimation of parameters. If we set $n_k = \sum_{i=1}^n \mathbf{1}\{y_i = k\}$, $n_{kj} = \sum_{i=1}^n \mathbf{1}(y_i = k, x_i^{(j)} = 1)$, the MLE can be written as

$$\hat{\pi}_k = \frac{n_k}{n}, \qquad \hat{\phi}_{kj} = \frac{\sum_{i:y_i=k} x_i^{(j)}}{n_k} = \frac{n_{kj}}{n_k}.$$

But the MLEs can be problematic in some cases. For example, if the ℓ-th word did not appear in any documents labelled as class k ($n_{kl} = 0$), then $\hat{\phi}_{k\ell} = 0$. But if we then wish to compute the predictive probability once for a new document $\tilde{x}$ which contains ℓ-th word, we have:

$$\begin{aligned} &p(\tilde{y} = k|\tilde{x} \text{ with } \ell\text{-th entry equal to } 1, \hat{\theta}) \\ &\propto \hat{\pi}_k \prod_{j=1}^{p} \left(\hat{\phi}_{kj}\right)^{\tilde{x}^{(j)}} \left(1 - \hat{\phi}_{kj}\right)^{1 - \tilde{x}^{(j)}} = 0, \end{aligned}$$

since $\hat{\phi}_{k\ell} = 0$. This means that we will never attribute a new document containing word ℓ to class k (regardless of what other words in it may be!). Moreover, probability of a document under all classes can be 0 by the same reasoning.

Let us consider a Bayesian approach to the same model. We can write the likelihood as

$$\begin{aligned} p(\mathcal{D}|\theta) = \prod_{i=1}^{n} p(x_i, y_i|\theta) &= \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\pi_k \prod_{j=1}^{p} \phi_{kj}^{x_i^{(j)}} (1 - \phi_{kj})^{1 - x_i^{(j)}} \right)^{\mathbf{1}(y_i = k)} \\ &= \prod_{k=1}^{K} \pi_k^{n_k} \prod_{j=1}^{p} \phi_{kj}^{n_{kj}} (1 - \phi_{kj})^{n_k - n_{kj}}. \end{aligned}$$

For a conjugate prior, we can use $\mathrm{Dir}((\alpha_k)_{k=1}^K)$ for π, and $\mathrm{Beta}(a, b)$ for ϕ_{kj} independently. Now, because the likelihood factorises, the posterior distribution over π and (ϕ_{kj}) also factorises, and posterior for π is $\mathrm{Dir}((\alpha_k + n_k)_{k=1}^K)$, and for ϕ_{kj} is $\mathrm{Beta}(a + n_{kj}, b + n_k - n_{kj})$. If we want to predict a label $\tilde{y}$ for a new document $\tilde{x}$, we obtain

$$p(\tilde{x}, \tilde{y} = k|\mathcal{D}) = p(\tilde{y} = k|\mathcal{D}) p(\tilde{x}|\tilde{y} = k, \mathcal{D})$$

with

$$\begin{aligned} p(\tilde{y} = k|\mathcal{D}) &= \frac{\alpha_k + n_k}{\sum_{l=1}^K \alpha_l + n} \\ p(\tilde{x}^{(j)} = 1|\tilde{y} = k, \mathcal{D}) &= \frac{a + n_{kj}}{a + b + n_k} \end{aligned}$$

and the predicted class is

$$p(\tilde{y}=k|\tilde{x},\mathcal{D}) = \frac{p(\tilde{y}=k|\mathcal{D})p(\tilde{x}|\tilde{y}=k,\mathcal{D})}{p(\tilde{x}|\mathcal{D})} \propto \frac{\alpha_k+n_k}{\sum_{l=1}^K \alpha_l + n} \prod_{j=1}^p \left(\frac{a+n_{kj}}{a+b+n_k}\right)^{\tilde{x}^{(j)}} \left(\frac{b+n_k-n_{kj}}{a+b+n_k}\right)^{1-\tilde{x}^{(j)}}.$$

Compared to the MLE plug-in predictions, pseudocounts help to "regularise" probabilities away from the extreme values.

7.3 Laplace Approximation

Bayesian approach to learning is conceptually very elegant, but the posterior distributions are intractable in almost all interesting cases, and we therefore need to resort to various approximations. One of the most popular techniques for approximation of intractable posterior distributions is the *Laplace* or *saddlepoint approximation*. The idea is to simply approximate the posterior distribution $p(\theta|\mathcal{D})$ with a (multivariate) Gaussian distribution. Given the ease of manipulating Gaussians, this is a convenient choice, since the various posterior expectations and predictive distributions will be easier to calculate when we have Gaussian approximate posteriors.

Consider for simplicity the case where parameter θ is a scalar and assume that posterior mode $\widehat{\theta}^{\text{MAP}}$ is available. Often, the posterior mode can be found even if the normalising constant $p(\mathcal{D})$ is intractable since it suffices to maximise $p(\theta|\mathcal{D}) \propto p(\theta,\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)$. Then, we can use a Taylor expansion of $\log p(\theta|\mathcal{D})$ around the mode $\widehat{\theta}^{\text{MAP}}$:

$$\begin{aligned}\log p(\theta|\mathcal{D}) &= \log p(\widehat{\theta}^{\text{MAP}}|\mathcal{D}) + \left.\frac{\partial \log p(\theta|\mathcal{D})}{\partial\theta}\right|_{\theta=\widehat{\theta}^{\text{MAP}}} \left(\theta-\widehat{\theta}^{\text{MAP}}\right) \\ &\quad + \left.\frac{\partial^2 \log p(\theta|\mathcal{D})}{\partial\theta^2}\right|_{\theta=\widehat{\theta}^{\text{MAP}}} \frac{\left(\theta-\widehat{\theta}^{\text{MAP}}\right)^2}{2} + \mathcal{O}\left(\left(\theta-\widehat{\theta}^{\text{MAP}}\right)^3\right).\end{aligned}$$

By ignoring the third and higher order terms and noticing that the the first derivative at the mode must be zero, we have an approximation:

$$\log p(\theta|\mathcal{D}) \approx \log p(\widehat{\theta}^{\text{MAP}}|\mathcal{D}) - \frac{\tau}{2}\left(\theta-\widehat{\theta}^{\text{MAP}}\right)^2, \tag{7.1}$$

where we write $\tau = -\frac{\partial^2 \log p(\theta|\mathcal{D})}{\partial\theta^2} \geq 0$. But recall that $\log \mathcal{N}\left(\theta|\mu,\sigma^2\right) = \log\left(\left(2\pi\sigma^2\right)^{-1/2}\right) - \frac{1}{2\sigma^2}\left(\theta-\mu\right)^2$, so this second order Taylor approximation has exactly the form of a normal log-density with mean $\mu = \widehat{\theta}^{\text{MAP}}$ and variance $\sigma^2 = \tau^{-1}$ so we can approximate the posterior with $\mathcal{N}\left(\widehat{\theta}^{\text{MAP}}, \left(-\frac{\partial^2 \log p(\theta|\mathcal{D})}{\partial\theta^2}\right)^{-1}\right)$.

This idea easily extends to multivariate densities. In particular, the Laplace approximation of $p(\theta|\mathcal{D})$ is a multivariate Gaussian $\mathcal{N}\left(\widehat{\theta}^{\text{MAP}}, \Sigma\right)$, where *the inverse covariance matrix is given by the negative Hessian of the log-posterior* evaluated at the posterior mode:

$$\Sigma^{-1} = \left.-\frac{\partial^2 \log p(\theta|\mathcal{D})}{\partial\theta\partial\theta^\top}\right|_{\theta=\widehat{\theta}^{\text{MAP}}}.$$

Since $\log p(\theta|\mathcal{D})$ agrees with $\log p(\theta, \mathcal{D})$ up to a constant, they have the same derivatives, so often we work with the *energy function* $J(\theta) = -\log p(\theta, \mathcal{D})$, which is the negative logarithm of the unnormalised posterior. Then we can write

$$\Sigma^{-1} = \left.\frac{\partial^2 J(\theta)}{\partial\theta\partial\theta^\top}\right|_{\theta=\widehat{\theta}^{\mathrm{MAP}}}.$$

7.4 Variational Bayes

Assume that we are taking a Bayesian approach to inference in a latent variable model $p(\mathbf{X}, \mathbf{z}|\theta)$ with observations $\mathbf{X}$, latent variables $\mathbf{z}$ and parameters θ. Now, our treatment of latent variables and parameters is exactly the same. We can now consider some joint distribution $q(\mathbf{z}, \theta)$ of latent variables and parameters, called variational distribution (like in EM, but note that it was not allowed to place a distribution over θ in EM!). We claim that the quantity

$$\mathcal{F}(q) = \mathbb{E}_q\left[\log p(\mathbf{X}, \mathbf{z}, \theta)\right] + H(q) \tag{7.2}$$

is a lower bound on log-evidence $\log p(\mathbf{X})$. We can write

$$\begin{aligned}
\mathcal{F}(q) &= \mathbb{E}_q\left[\log p(\mathbf{X}, \mathbf{z}, \theta)\right] - \mathbb{E}_q[\log q(\mathbf{z}, \theta)] \\
&= \mathbb{E}_q\left[\log p(\mathbf{z}, \theta|\mathbf{X})\right] + \log p(\mathbf{X}) - \mathbb{E}_q[\log q(\mathbf{z}, \theta)] \\
&= -\mathrm{KL}\left(q(\mathbf{z}, \theta)||p(\mathbf{z}, \theta|\mathbf{X})\right) + \log p(\mathbf{X}),
\end{aligned}$$

which is by Gibbs inequality maximised (and equal to log-evidence) when KL is zero, i.e. when $q(\mathbf{z}, \theta) = p(\mathbf{z}, \theta|\mathbf{X})$. Thus, for any variational distribution q, $\mathcal{F}(q) \leq \log p(x)$. Expression 7.2 is called the *evidence lower bound (ELBO)*.

To reason about all the unknowns in the model, we would simply need to compute the joint posterior $p(\mathbf{z}, \theta|\mathbf{X})$, but this is almost always intractable. Hence, the variational Bayesian inference *approximates* the posterior by starting with a family $\mathcal{Q}$ of tractable variational distributions $q(\mathbf{z}, \theta)$ (e.g. $q(\mathbf{z}, \theta|\eta)$ where η are the *variational parameters*), and aims to minimise the divergence $\mathrm{KL}\left(q(\mathbf{z}, \theta)||p(\mathbf{z}, \theta|\mathbf{X})\right)$ over $\mathcal{Q}$ or, equivalently, maximise the ELBO, i.e. find the tightest lower bound on the log-evidence.

Consider family of variational distributions which factorise: $q(\mathbf{z}, \theta) = q_{\mathbf{Z}}(\mathbf{z})\, q_\Theta(\theta)$. For a fixed q_Θ, we can solve for $q_{\mathbf{Z}}$ which maximises ELBO (*exercise):*

$$q_{\mathbf{Z}}(\mathbf{z}) \propto \exp\left(\int \log p(\mathbf{X}, \mathbf{z}, \theta) q_\Theta(\theta)\, d\theta\right),$$

and by symmetry, for a fixed $q_{\mathbf{Z}}$, we can solve for q_Θ which maximises ELBO:

$$q_\Theta(\theta) \propto \exp\left(\int \log p(\mathbf{X}, \mathbf{z}, \theta) q_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z}\right).$$

Now, one can formulate an algorithm similar to EM, which alternates between optimising $q_{\mathbf{Z}}$ and q_Θ, such that each iteration increases ELBO and thus decreases the KL divergence from the posterior.

7.5 Bayesian Model Selection

Consider a situation where we do not have one Bayesian model but several. Each model $\mathcal{M}$ has a set of parameters $\theta_{\mathcal{M}}$, likelihood $p(\mathcal{D}|\theta_{\mathcal{M}})$ and the prior distribution $p(\theta_{\mathcal{M}})$. Within each model, the posterior distribution is

$$p(\theta_{\mathcal{M}}|\mathcal{D},\mathcal{M}) = \frac{p(\mathcal{D}|\theta_{\mathcal{M}},\mathcal{M})p(\theta_{\mathcal{M}}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})}$$

where the normalising constant is the marginal probability of the data under model $\mathcal{M}$ (*Bayesian model evidence*):

$$p(\mathcal{D}|\mathcal{M}) = \int_{\Theta} p(\mathcal{D}|\theta_{\mathcal{M}},\mathcal{M})p(\theta_{\mathcal{M}}|\mathcal{M})d\theta$$

In Bayesian model selection, one compares models using their *Bayes factors* $\frac{p(\mathcal{D}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M}')}$.

Considering Bayesian model evidence can be interpreted as a Bayesian version of *Occam's Razor*: of two explanations adequate to explain the same set of observations, the simpler should be preferred. Namely, note that the model evidence $p(\mathcal{D}|\mathcal{M})$ is the probability that a set of randomly selected parameter values (under the prior) inside the model would generate dataset $\mathcal{D}$. In that case, models that are *too simple* are unlikely to generate the observed dataset. On the other hand, models that are *too complex* can generate many possible datasets, so again, they are unlikely to generate that particular dataset at random.