1 Generative models

1.1 Generative Models for Discrete Data

1.1.1 Bayesian Concept Learning

can learn a concept $c \in C$ from positive examples alone. For that define the posterior: $p(c|\mathcal{D})$. To get to learn a concept you need a hypothesis space $\mathcal{H}$ and a version space (a subset of $\mathcal{H}$) that is consistent with $\mathcal{D}$. You choose a hypothesis h by assuming that samples are randomly chosen from the true concept and calculate $p(\mathcal{D}|h) = [\frac{1}{|h|}]^N$ (sampling the N data points from h). You than choose the hypothesis that has the highest probability (thereby you choose suspicious coincidences of too broad models). The priors can be chosen e.g. by giving lower priority to concepts with complex rules (e.g. "all powers of 2 below 100 but not 64."). This is subjective, however often beneficial for rapid learning. Using Bayes rule, we can calculate the posterior:

$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h)p(h)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|h)p(h)}{\sum_{h' \in \mathcal{H}} p(\mathcal{D}|h')p(h')} = \frac{\mathbb{I}(\mathcal{D} \in h)p(h)}{\sum_{h' \in \mathcal{H}} \mathbb{I}(\mathcal{D} \in h')p(h')},$$

where $\mathbb{I}(\mathcal{D} \in h)p(h) = 1$ if the data adhere to the h. The maximum of $p(h|\mathcal{D})$ is the **MAP estimate**.

With more data the MAP-estimate converges to the MLE. If the true hypothesis is in $\mathcal{H}$ then MLE and MAP will converge to it (→ consistent estimators). If you take the entire distribution of the hypotheses you get a distribution for the estimate (and not a point prediction) → **posterior predictive distribution**.

$$p(\tilde{x}|\mathcal{D}) = \sum_h p(\tilde{x}|h)p(h|\mathcal{D})$$

This weighting of hypotheses is called **Bayes model averaging**. For small data sets you get a vague posterior and broad predictive distribution. You can replace the posteriors with their delta-function:

$$p(\tilde{x}|\mathcal{D}) = \sum_h p(\tilde{x}|h)\delta_{\hat{h}_{\mathsf{MAP}}}(h)$$

→ **plug-in approximation** (under-represents uncertainty).

1.1.2 Beta-binomial model

This is a distribution that uses a binomial distribution as its likelihood and a beta-distribution over it's θ parameter as its prior.

Likelihood

$$p(\mathcal{D}|\theta) = \mathrm{Bin}(k|n,\theta) \propto \theta^k(1-\theta)^{n-k},$$

where k are the successful trials, n are the total trials and θ are the success-probabilities of the single experiments. k and $n-k$ are *sufficient statistics of the data*: $p(\theta|\mathcal{D} = p(\theta|k, n-k))$.

Prior

$$\mathrm{Beta}(\theta|\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

the parameters α and β are used as hyper parameters. The prior has the same form as the likelihood → *conjugate prior*.

Posterior

$$p(\theta|\mathcal{D}) \propto \mathsf{Bin}(k|n,\theta)\mathsf{Beta}(\theta|\alpha,\beta) = \mathsf{Beta}(\theta|k+\alpha, n-k+\beta)$$

Q → We add pseudo-counts (α, β) to empirical counts (N, k). The posterior predictive distribution is:

$$\begin{aligned} p(\tilde{x}=1|\mathcal{D}) &= \int_0^1 p(\tilde{x}=1|\theta)p(\theta|\mathcal{D})\mathrm{d}\theta = \int_0^1 \theta\mathsf{Beta}(\theta|N-k+\alpha, k+\beta)\mathrm{d}\theta \\ &= \mathsf{E}[\theta|\mathcal{D}] = \frac{N-k+\alpha}{N\cancel{-k+k}+\alpha+\beta} \end{aligned}$$

! → If we used the MLE for θ instead $(p(\theta|\mathcal{D}) = p(\theta|\theta_{\mathsf{MLE}}))$ and we only had little data and no failures (e.g. 3 coin flips and all are tails). Then the MLE estimate would be $\theta_{\mathsf{MLE}}) = 0/3 = 0$. This is called *zero count estimate* problem or *black swan paradox* (i.e. you don't attribute possibilities to something you have never seen before). Solution: You use a uniform prior $(\alpha = \beta = 1)$: $p(\tilde{x}=1|\mathcal{D}) = \dfrac{N-k+1}{N+2}$.

1.1.3 Dirichlet-multinomial model

Before: model of k successes, now: k times a discrete outcome (e.g. four pips on a die roll) in n experiments.

Prior: The Dirichlet distribution:

$$\mathsf{Dir}(\theta|\alpha) = \frac{1}{B(\alpha)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}$$

Posterior:

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto p(\mathcal{D}|\theta)p(\theta|\alpha) \\ &\propto \prod_{k=1}^{K}\theta_k^{N_k}\theta_k^{\alpha_k-1} = \prod_{k=1}^{K}\theta_k^{N_k+\alpha_k-1} \\ &= \mathsf{Dir}(\theta|\alpha_1+N_1, \cdots, \alpha_K+N_K) \end{aligned}$$

The posterior predictive distribution is:

$$\begin{aligned} p(\tilde{\tilde{X}}=j|\mathcal{D}) &= \int p(\tilde{X}=j|\theta)p(\theta|\mathcal{D})\mathrm{d}\theta \\ &= \int p(\tilde{X}=j|\theta_j)\left[\int p(\theta_{-j},\theta_j|\mathcal{D})\mathrm{d}\theta_{-j}\right]\mathrm{d}\theta_j \\ &= \int \theta_j p(\theta_j|\mathcal{D})\mathrm{d}\theta_j = \mathsf{E}[\theta_j|\mathcal{D}] = \frac{N_j+\alpha_j}{\sum_k N_k+\alpha_k} \end{aligned}$$