Active Learning

Laura, Peter, Simon

General introduction

Notation:

Data $\mathcal{D}$. New data $\mathcal{D}^*$. Parameters θ. Input $\mathbf{x}$. Output $\mathbf{y}$. Model $\mathcal{M}$.

NB: We might not mention the $\mathcal{M}$ explicitly in all the probabilities.

We are interested in the posterior probability of the model paramters given the data

$$P(\theta|\mathcal{D},\mathcal{M}) = \frac{P(\mathcal{D}|\theta,\mathcal{M})\pi(\theta,\mathcal{M})}{P(\mathcal{D},\mathcal{M})} \tag{1}$$

The *predictive distribution* is the probability $P(\mathcal{D}^*|\mathcal{D},\mathcal{M})$ of observing a new data point $\mathcal{D}^*$ given the old data $\mathcal{D}$ and a model $\mathcal{M}$.

$$\begin{aligned}
P(\mathcal{D}^*|\mathcal{D},\mathcal{M}) &= \int \mathrm{d}\theta P(\mathcal{D}^*,\theta|\mathcal{D},\mathcal{M}) \\
&= \int \mathrm{d}\theta P(\mathcal{D}^*|\mathcal{D},\mathcal{M},\theta)P(\theta|\mathcal{D},\mathcal{M}) \\
&= \int \mathrm{d}\theta P(\mathcal{D}^*|\mathcal{M},\theta)P(\theta|\mathcal{D},\mathcal{M}) \,,
\end{aligned}$$

where in the last step we used $P(\mathcal{D}^*|\mathcal{D},\mathcal{M},\theta) = P(\mathcal{D}^*|\mathcal{M},\theta)$, because the new data should depend only on the model and the parameters and not on the collected data. That is, we assume that the model captures all the structure in the data. This assumption is typical for Bayesian inference.

We want to maximize the expected information gain of the input $\mathbf{x}$:

$$\mathcal{U}(\mathbf{x}) = H[p(\theta|\,\mathcal{D})] - \mathbb{E}_{P(\mathbf{y}|\mathbf{x},\,\mathcal{D})} H[P(\theta|\,\mathcal{D}, \mathbf{x}, \mathbf{y})]\,, \tag{2}$$

which corresponds to minimizing the second term. This is called posterior entropy minimization. We can get an alternative formulation by noting that

$$\begin{aligned}\mathcal{U}(\mathbf{x}) &= I[\theta, \mathbf{y}|\,\mathcal{D}, \mathbf{x}] \qquad (3)\\ &= H[P(\mathbf{y}|\mathbf{x},\,\mathcal{D})] - \mathbb{E}_{P(\theta|\,\mathcal{D})} H[P(\mathbf{y}|\mathbf{x}, \theta)]\,, \qquad (4)\end{aligned}$$

where I is the mutual information, which is symmetric in its arguments. Writing it this way allows for a different interpretation of $\mathcal{U}(\mathbf{x})$ Now $H[P(\mathbf{y}|\mathbf{x},\,\mathcal{D})]$ should be large, which makes sense, because we should choose an input $\mathbf{x}$ for which we don't know yet what the output $\mathbf{y}$ will be. Furthermore $\mathbb{E}_{P(\theta|\,\mathcal{D})} H[P(\mathbf{x}|\mathbf{y}, \theta)]$ should be small, because we don't want to choose an input $\mathbf{x}$ for which the output $\mathbf{y}$ is very uncertain. NB: Copied from Houlsby thesis: In other words, we seek the input $\mathbf{x}$ for which the parameters under the posterior make confident predictions (term 2), but these predictions are highly diverse. That is, the parameters disagree about the output $\mathbf{y}$, hence this formulation is named Bayesian Active Learning by Disagreement (BALD).

1 Experiment

We show two gratings to participants. The frequency of the grating is fixed. The orientation of both gratings is the same but varied in each trial to avoid afterimages. One grating is always of the same contrast level, but the side is chosen at random.

We vary the contrast of the other grating. We characterize the difference in contrast between the grating that is shown on the left and the grating that is shown on the right with x. For negative x the grating on the left is of higher contrast, for positive x the stimulus on the right is of higher contrast. If we denote the fixed baseline contrast with x_b, then the value of x is in the range $[-(1-x_b),(1-x_b)]$.

We chose the presented x according to different strategies and record the answers (left L, or right R) of the participants when the decide on which side the grating with higher contrast is shown.

1.1 Choosing the presented stimulus

We assume a prior $\pi(\theta)$ where $\theta = \{w_0, w_1, \lambda\}$ is a set of parameters that describe our model. As the model $\mathcal{M}$ we use a sigmoid function

$$\sigma\left(\theta, x\right) = \lambda/2 + \frac{1-\lambda/2}{1+\exp[-w_1(x-w_0)]}\,, \tag{5}$$

where λ is the lapse rate. The lapse rate accounts for wrong answers that are not because the task was to difficult, but because the participant hit mistakenly hits the wrong button. NB: We drop the model $\mathcal{M}$ in all the subsequent probabilities. Whenever we use θ we mean the parameters together with the sigmoid model. We collect data N data points by presenting a stimulus $x \in [-1.0, 1.0]$ and observing a binary response $y \in \{0,1\}$:

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\} \equiv (X^N, Y^N) \tag{6}$$

NB: We drop the N, if it is not needed to dissociated the steps. The likelihood of the parameters θ given the data D is given by

$$P(Y^N|\theta, X^N) = \prod_{i=1}^{N} P(y_i|\theta, x_i) \tag{7}$$

$$= \prod_{i=1}^{N} \sigma(\theta, x)^{y_i} (1 - \sigma(\theta, x))^{1-y_i} \tag{8}$$

NB: The stimuli x are not considered part of the data, because we have control over it The posterior probability of the parameters $\mathbf{w}$ is

$$P(\theta|X, Y) = \frac{P(Y|\theta, X)\pi(\theta)}{P(Y|X)} \tag{9}$$

The denominator, i.e. the marginal likelihood, is computed by taking the integral over all hypotheses:

$$P(Y|X) = \int P(Y|\theta', X)\pi(\theta')\,\mathrm{d}\theta' \tag{10}$$

The goal is to get a posterior $P(\mathbf{w}|X, Y)$ that is of low uncertainty. We use entropy as a measure of the current uncertainty of our estimation of θ. By *current* we mean that we use the data $\mathcal{D}$ we have discovered in the N steps until now. The new data points are labeled x, y.

$$H[P(\theta|\,\mathcal{D})] = -\int P(\theta'|\,\mathcal{D}) \log[P(\theta'|\,\mathcal{D})]\,\mathrm{d}\theta' \,. \tag{11}$$

In principle we would now like to choose our next stimulus x such that it minimizes the resulting entropy $H(\mathbf{w}|X_N, Y_N, x, y)$, but we do not know what y is going to be.

So we want to find the x that minimizes the mean:

$$H[P(\mathbf{w}|\,\mathcal{D},x,y=0)]\;P(y=0|\,\mathcal{D},x) \tag{12}$$

$$+\,H[P(\mathbf{w}|\,\mathcal{D},x,y=1)]\;P(y=1|\,\mathcal{D},x)\,. \tag{13}$$

Here $P(y=0/1|\,\mathcal{D},x)$ is called the *predictive distribution*. Determining them again requires an integral over the hypotheses:

$$P(y=0/1|\,\mathcal{D},x)=\int P(y=0/1|\theta,x)P(\theta|\,\mathcal{D})\,\mathrm{d}\theta \tag{14}$$

NB: The above is the direct approach without using BALD learning.

Instead we should use BALD learning and find the x that maximizes:

$$\mathcal{U}(x)=H[P(y|\,\mathcal{D},x)]-\mathbb{E}_{P(\theta|\,\mathcal{D})}H[P(y|\theta,x)]\,. \tag{15}$$

1.2 Humans

Here the contrast difference x is chosen by a human. They can select every possible value for x. To help them they

2 Approximations

We need to determine $K(x)$ for all x values that we consider as worthwhile new stimuli. This can be many values and doing the involved integrals over posteriors is costly. There are several ways to deal with this problem.

2.1 Restricting the tested stimuli

We need to discretize the x anyways. We choose values ....

2.2 Approximating the posterior

We often determine the mean of a function over the posterior distribution. This is computationally expensive, in particular for large parameter spaces. If we take samples of the posterior and approximate the integrals by smaller sums over the samples, we can save computation time. To get good samples from the posterior distribution we use the Metropolis-Hastings algorithm as an implementation of Markov Chain Monte Carlo integration. As a proposal distribution $Q(x; x')$ we choose a multi variate normal distribution which determines the random walk that samples from the posterior.

2.3 Focused active learning

Instead of maximizing the utility function $\mathcal{U}(x)$ over can maximize it only with respect to a subset of parameters. We In the scenario of the sigmoid we might be interested in choosing the x where the entropy is w_1 that maximizes