Active Learning

Laura, Peter, Simon

General introduction

Data $\mathcal{D}$. Parameters θ. Input $\mathbf{x}$. Output $\mathbf{y}$. We want to maximize the expected information gain of the input $\mathbf{x}$:

$$\mathcal{U}(\mathbf{x}) = H[p(\theta|\,\mathcal{D})] - \mathbb{E}_{P(\mathbf{y}|\mathbf{x},\,\mathcal{D})} H[P(\theta|\,\mathcal{D}, \mathbf{x}, \mathbf{y})]\,, \tag{1}$$

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

$$\begin{aligned} \mathcal{U}(\mathbf{x}) &= I[\theta, \mathbf{y}|\,\mathcal{D}, \mathbf{x}] &(2)\\ &= H[P(\mathbf{y}|\mathbf{x},\,\mathcal{D})] - \mathbb{E}_{P(\theta|\,\mathcal{D})} H[P(\mathbf{y}|\mathbf{x}, \theta)]\,, &(3) \end{aligned}$$

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).

NB: Maybe start even earlier, with the most important points of the Houlsby introduction.

We assume a prior $\pi(w)$ where $\mathbf{w}$ is a vector of parameters that describe our model. As the model we use a sigmoid function

$$\sigma\left(w_0, w_1, x\right) = \frac{1}{1 + \exp[-(w_0 - w_1 x)]} \tag{4}$$

NB: This might not be the best way to write the sigmoid. We collect data N data points by presenting a stimulus $x \in \mathbb{R}$ and observing a binary response $y \in \{0, 1\}$:

$$D = \{(x_1, y_1), \ldots, (x_N, y_N)\} \equiv (X, Y) \tag{5}$$

The likelihood of the parameters $\mathbf{w}$ given the data D is given by

$$\begin{aligned} P(Y \mid \mathbf{w}, X) &= \prod_{i=1}^{n} P(y_i \mid \mathbf{w}, x_i) && (6) \\ &= \prod_{i=1}^{n} \sigma\left(w_0 + w_1 x\right)^{y_i} \left(1 - \sigma\left(w_0 + w_1 x\right)\right)^{1-y_i} && (7) \end{aligned}$$

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

$$P(\mathbf{w} \mid X, Y) = \frac{P(Y \mid w, X)\pi(\mathbf{w})}{P(Y \mid X)} \tag{8}$$

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

$$\int P(Y \mid \mathbf{w}', X)\pi(\mathbf{w}')\,\mathrm{d}\mathbf{w}' = \iint P(Y \mid w_0, w_1, X)\pi(w_0, w_1)\,\mathrm{d}w_0'\,\mathrm{d}w_1' \tag{9}$$

The goal is to get a posterior $P(\mathbf{w} \mid X, Y)$ that is of low uncertainty. We use entropy as a measure of the current uncertainty of our estimation of $\mathbf{w}$. By *current* we mean that we use the data we have discovered in the n steps until now. To make this clear we write X_N, Y_N instead of X, Y:

$$H[P(\mathbf{w} \mid X_N Y_N)] = -\int P(\mathbf{w}' \mid X_N, Y_N)\log[P(\mathbf{w}' \mid X_N, Y_N)]\,\mathrm{d}\mathbf{w}'\,. \tag{10}$$

In principle we would now like to choose our next stimulus x_{N+1} such that it minimizes the resulting entropy $H(\mathbf{w} \mid X_N, Y_N, x_{N+1}, y_{N+1})$, but we do not know what y_{N+1} is. So we want to find the x_{N+1} that minimizes the mean:

$$K(x_{N+1}) = H[P(\mathbf{w} \mid X_N, Y_N, x_{N+1}, y_{N+1} = 0)]\; P(y_{N+1} = 0 \mid X_N, Y_N, x_{N+1}) \tag{11}$$

$$+ H[P(\mathbf{w} \mid X_N, Y_N, x_{N+1}, y_{N+1} = 1)]\; P(y_{N+1} = 1 \mid X_N, Y_N, x_{N+1})\,. \tag{12}$$

Here $P(y_{N+1} = 0/1 \mid X_N, Y_N, x_{N+1})$ is called the *predictive distribution*. Determining them again requires an integral over the hypotheses:

$$P(y_{N+1} = 0/1 \mid X_N, Y_N, x_{N+1}) = \int P(y_{N+1} = 0/1 \mid \mathbf{w}', x_{N+1})P(\mathbf{w}' \mid X_N, Y_N)\,\mathrm{d}\mathbf{w}' \tag{13}$$

17 For