Active Learning – CCCN 2016, Lisbon

Laura Driscoll, Simon Weber, Peter Zatka-Haas

September 7, 2016

General introduction

0.1 Uncertainties in determining model parameters from data

- **Parameter uncertainty:** Observed data sets are finite. Therefore it is impossible to determine the parameters of a model precisely.
  In our example we can't be confident about the psychometric curve if we have few data points.

- **Inherent uncertainty:** Typically there is noise in observing the data. In other words, there is randomness in the data that is not explained by the model. Inherent uncertainty is also called **observation noise**.
  In our example the inherent uncertainty is high at low contrast differences.

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 parameters given the data

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

The posterior reflects the parameters for which the model best describes the data and the underlying uncertainty (parameter uncertainty). Another probability that occurs frequently in the following is the *predictive distribution*. 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 $\mathcal{M}$.

$$P(\mathcal{D}^*|\mathcal{D},\mathcal{M}) = \int \mathrm{d}\theta P(\mathcal{D}^*,\theta|\mathcal{D},\mathcal{M}) \tag{2}$$

$$= \int \mathrm{d}\theta P(\mathcal{D}^*|\mathcal{D},\theta,\mathcal{M}) P(\theta|\mathcal{D},\mathcal{M}) \tag{3}$$

$$= \int \mathrm{d}\theta \underbrace{P(\mathcal{D}^*|\theta,\mathcal{M})}_{\text{inherent uncertainty}} \underbrace{P(\theta|\mathcal{D},\mathcal{M})}_{\text{parameter uncertainty}}, \tag{4}$$

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. Note how Eq. (4) contains both aforementioned uncertainties.

It is typically desirable to have little uncertainty in the posterior. The amount of uncertainty can be measured by the entropy

$$H[P(\theta|\mathcal{D})] = -\mathbb{E}_{\theta\sim P(\theta|\mathcal{D})} \log P(\theta|\mathcal{D}) \tag{5}$$

In the psychophysics experiment that we conduct, the data is comprised of stimulus-response pairs. We can choose the stimulus $\mathbf{x}$ and observe the response $\mathbf{y}$; see Fig. 1 for the graphical model. Say we have collected some stimulus-response pairs already (represented with $\mathcal{D}$). The goal is to choose the next stimulus $\mathbf{x}$ such that the uncertainty in the posterior is decreased. In other words, we would like to

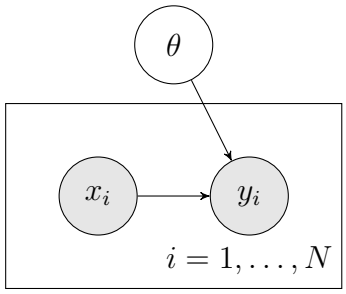

Figure 1: Graphical model. White notes indicate latent (unobserved) variables and shaded notes denote observed variables. The stimulus x is independent of the parameters θ. Adapted from Houlsby thesis.

choose $\mathbf{x}$ such that the corresponding decrease in entropy

$$H[p(\theta|\,\mathcal{D})] - H[p(\theta|\,\mathcal{D}, \mathbf{x}, \mathbf{y})] \tag{6}$$

is maximal. But we don't know the answer $\mathbf{y}$ that we are going to get. We can only maximize the decrease in expected posterior entropy:

$$\mathcal{U}(\mathbf{x}) = H[p(\theta|\,\mathcal{D})] - \mathbb{E}_{\mathbf{y}\sim P(\mathbf{y}|\mathbf{x},\,\mathcal{D})} H[P(\theta|\,\mathcal{D}, \mathbf{x}, \mathbf{y})]\,, \tag{7}$$

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

$$\mathcal{U}(\mathbf{x}) = I[\theta, \mathbf{y}|\,\mathcal{D}, \mathbf{x}] \tag{8}$$

$$= \underbrace{H[P(\mathbf{y}|\mathbf{x},\,\mathcal{D})]}_{\text{marginal entropy}} - \underbrace{\mathbb{E}_{\theta\sim P(\theta|\,\mathcal{D})} H[P(\mathbf{y}|\mathbf{x}, \theta)]}_{\text{expected conditional entropy}}\,, \tag{9}$$

where I is the mutual information, which is symmetric in its arguments. Writing it this way allows for a different interpretation of the utility function $\mathcal{U}(\mathbf{x})$. The first

term (the marginal entropy) is the entropy of the predictive distribution:

$$H\left[P(\mathbf{y}|\mathbf{x},\,\mathcal{D})\right] = H\left[\int \mathrm{d}\theta' P(\mathbf{y}|\mathbf{x},\theta')P(\theta'|\,\mathcal{D})\right] . \tag{10}$$

This term reflects the parameter uncertainty. It should be large, because we want to choose an input $\mathbf{x}$ for which the parameters θ' disagree about the output $\mathbf{y}$. If they would all agree, we would not get useful information about which parameters are better than others. More precisely, if for different θ' different $\mathbf{y}$ are likely, than the integral results in a non-vanishing probability for many values of $\mathbf{y}$. We thus get a high entropy.

The second term (the expected conditional entropy) is:

$$\mathbb{E}_{\theta\sim P(\theta|\,\mathcal{D})}H[P(\mathbf{y}|\mathbf{x},\theta)] = \int \mathrm{d}\theta' P(\theta'|\,\mathcal{D})H\left[P(\mathbf{y}|\mathbf{x},\theta')\right] . \tag{11}$$

This term reflects the inherent uncertainty because it is high if all $\mathbf{y}$ are equally probable for the given parameter θ'. The term should be small. This makes sense, because if the prediction of the output $\mathbf{y}$ given by the model parameters θ' is very uncertain at the input $\mathbf{x}$, we learn very little about the choice of the parameters after we observed the response. We thus should choose an $\mathbf{x}$ for which the parameter sets θ' typically make a confident prediction of the response. Now $H[P(\mathbf{y}|\mathbf{x},\,\mathcal{D})]$ should be large, which makes sense, because we should choose an input. 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 [term 1]. That [term 1] is, the parameters disagree about the output $\mathbf{y}$, hence this formulation is named Bayesian Active Learning by Disagreement (BALD). The different width of the expected conditional entropy and the marginal entropy as a

Figure 2: MISSING

function of the stimulus explain the typically Mexican hat like shape of the utility function $\mathcal{U}(\mathbf{x})$ in our experiments; see Fig. 2.

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 $P(\theta)$ 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

$$\sigma(\theta, x) = \lambda/2 + \frac{1-\lambda/2}{1+\exp[-w_1(x-w_0)]}, \tag{12}$$

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 $y \in \{0, 1\}$:

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\} \equiv (X^N, Y^N) \tag{13}$$

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

$$P(Y^N|\theta, X^N) = \prod_{i=1}^{N} P(y_i|\theta, x_i) \tag{14}$$

$$= \prod_{i=1}^{N} \sigma(\theta, x)^{y_i} (1 - \sigma(\theta, x))^{1-y_i} \tag{15}$$

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

$$P(\theta|X, Y) = \frac{P(Y|\theta, X)P(\theta)}{P(Y|X)} \tag{16}$$

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

$$P(Y|X) = \int P(Y|\theta', X)P(\theta') \, \mathrm{d}\theta' \tag{17}$$

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.

$$H[P(\theta|\,\mathcal{D})] = -\int P(\theta'|\,\mathcal{D}) \log[P(\theta'|\,\mathcal{D})]\,\mathrm{d}\theta'\,. \tag{18}$$

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:

$$H[P(\mathbf{w}|\,\mathcal{D}, x, y=0)]\;P(y=0|\,\mathcal{D}, x) \tag{19}$$

$$+\,H[P(\mathbf{w}|\,\mathcal{D}, x, y=1)]\;P(y=1|\,\mathcal{D}, x)\,. \tag{20}$$

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

$$P(y=0/1|\,\mathcal{D}, x) = \int P(y=0/1|\theta, x) P(\theta|\,\mathcal{D})\,\mathrm{d}\theta \tag{21}$$

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

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

$$\mathcal{U}(x) = H[P(y|\,\mathcal{D}, x)] - \mathbb{E}_{\theta\sim P(\theta|\,\mathcal{D})} H[P(y|\theta, x)]\,. \tag{22}$$

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

3 Paradox of ladder stimulus presentation

Assume a stimulus range from $-a$ to a. Imagine we present stimuli in the following way:

- We present the first at the center ($x_1 = 0$) and record the answer y_1
- For the n-th stimulus ($n > 1$) we take

$$x_n = x_{n-1} + (-1)^{y_{n-1}} \frac{a}{2^{n-1}} \tag{23}$$

For $N \to \infty$, this leads to a distribution of data points that seems to be best fitted by a step function. When we collect data this way, we know that it will always lead to a distribution that looks this way and we know that we should not conclude that the underlying psychometric function should have a steep step. If, however, someone would present us this data and claim that it was obtained without using the ladder algorithm above, we might be tempted to conclude that the underlying psychometric curve is very steep. This is paradoxical, because the likelihood principle claims that the data is all we need to determine our parameter. It should not matter how the data was obtained.

The solution of this paradox is that we should not conclude a step function in neither of the two cases. Instead the posterior suggests that all sigmoids that cross

$y^* = \frac{\#\text{ of response 1 around } x_\infty}{\#\text{ of response 0 around } x_\infty}$ at the stimulus value x_∞ to which the ladder method converges are equally probable. NB: 'around' is not properly defined here, but it can't be x_∞ precisely, because it is never reached. It should be some ε-ball. With our parameterization this is the set of sigmoids for which

$$w_0 = -x_\infty w_1 + c\,, \tag{24}$$

where $c > 0$ if $y^* > 1/2$, $c = 0$ if $y^* = 1/2$ and $c < 0$ if $y^* < 1/2$. So for $N \to \infty$ the posterior $P(w_0, w_1|\,\mathcal{D})$ turns into a line that is characterized by Eq. (24).

Marginalizing over the threshold value w_0 leads to a flat posterior for the slope $P(w_1|\,\mathcal{D})$. This shows that all slopes are equally likely and Bayesian inference does not conclude a step function.