Advanced Machine Learning (GR5242)
Fall 2019

Homework 3

Due: Wednesday 23 October, at 4pm, for both sections of the class

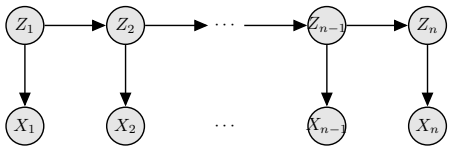
Homework submission: Please submit your homework by publishing a notebook that cleanly displays your code, results, and plots to pdf or html.

Problem 1: Short questions

(a) Suppose you have an algorithm that, when run n times, generates n independent samples from a density p. How do you generate n independent points distributed uniformly on the area under the curve p?

(b) A logistic regression classifier $\sigma(\mathbf{v}^t\mathbf{x} - c)$ is trained by fitting the vector $\mathbf{v}$ and offset c using an optimization algorithm. How does overfitting occur?

(c) Suppose the variables $Z_1, \ldots, Z_n$ in the graphical model below are unobserved. Are X_n and X_2 dependent or independent? Please explain your answer.



Problem 2: XOR network

Consider a neural network which:

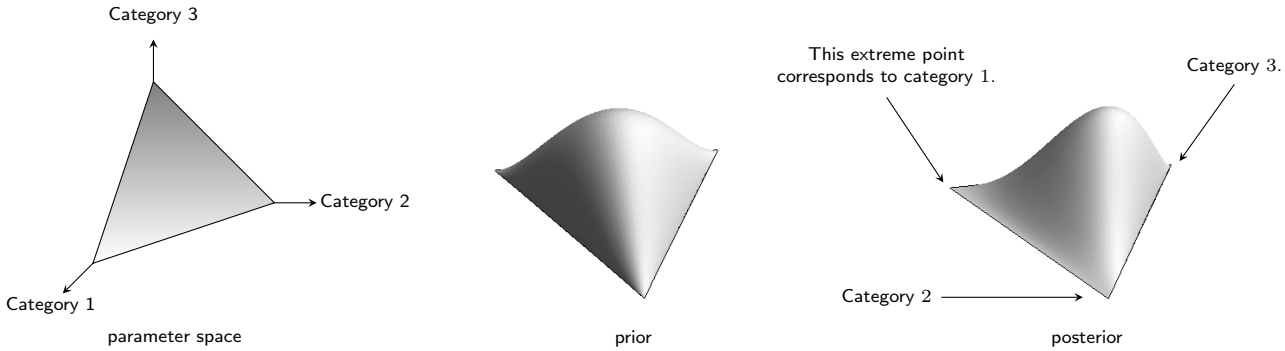
- Takes inputs in $\mathbf{x} = (x_1, x_2) \in \{-1, 1\}^2$.
- Has a single hidden layer with two vertices.
- Represents the function

$$f(\mathbf{x}) := \begin{cases} 1 & \text{if } x_1 = x_2 \\ 0 & \text{otherwise} \end{cases}$$

There several possible networks with these properties. Please draw such a network such that all weights and biases (=constant input units, if you need any) only take values 1 or -1.

Problem 3: Dirichlet distributions

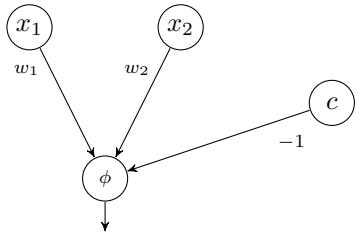
Consider a multinomial distribution on the set of categories $\mathbf{X} = \{1, 2, 3\}$. Recall that the parameter space of this distribution (the set of all vectors $\theta = (\theta_1, \theta_2, \theta_3)$ with non-negative entries and $\theta_1 + \theta_2 + \theta_3 = 1$) can be plotted as the area within a triangle, with each corner corresponding to one category (left figure):



The plot in the middle shows the density of a Dirichlet prior on the parameter space of the multinomial; the plot on the right is the resulting posterior given observed data. Does the observed sample in this case consist of one data point in category 3, or of one observation each in category 1 and 2? Please explain your answer.

Problem 4: Neural network classifier

Consider the classifier on $\mathbb{R}^2$, given by the following neural network with $\phi(x) := \mathbb{I}\{x \geq 0\}$:



Suppose that

$$\mathbf{w} := \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \qquad c := \frac{1}{2\sqrt{2}} \qquad \mathbf{x} := \begin{pmatrix} -3 \\ 0 \end{pmatrix} \qquad \mathbf{x}' := \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} .$$

Compute the classification result for $\mathbf{x}$ and for $\mathbf{x}'$.

Problem 5: Rejection and importance sampling

Suppose that we need samples from a density on $[0, 1]$, given by:

$$p(x) = 2x \ .$$

We want to estimate the mean of p using rejection sampling, and use the uniform distribution q on $[0, 1]$ as proposal distribution. We scale p to

$$\tilde{p} = \frac{1}{M} p$$

so that its density does not exceed q.

(a) What is the optimal choice for M if we wish to keep the number of rejected samples as small as possible?

(b) Compute the acceptance probability for the rejection sampling procedure (for the optimal choice of M). If we wish to obtain m samples from the rejection sampling procedure, how many times n do we have to sample from the proposal distribution on average?

(c) Now suppose we use importance sampling rather than rejection sampling, again with uniform proposal distribution. What is the variance of the importance sampling estimate of the mean, for n samples?

Problem 6: Rejection sampling without rejection

Let $p(\cdot)$ be a density function on $\mathbb{R}$, and suppose we want to estimate

$$\theta := \mathbb{E}_{X\sim p}\left[h(X)\right] = \int h(x)p(x)\,dx$$

One way of doing so is to draw samples from p using a rejection sampler given a proposal density $r(\cdot)$. We assume that we know a k such that $p(x) \leq kr(x)$ for all $x \in \mathbb{R}$. In this question we will assume that $k > 1$.

Consider the following variant of rejection sampling, where we instead keep all of the samples and label whether we would have accepted or rejected them in a standard rejection sampler. That is, for $i = 1, \ldots, N$:

- Sample $X_i \sim p(\cdot)$ and $U_i \sim \mathrm{Uniform}[0,1]$ independently.
- If $U_i \leq \frac{p(X_i)}{kr(X_i)}$, set a label $Z_i = A$ (**A**ccept); otherwise set $Z_i = R$ (**R**eject).

Suppose all the labels are not either all A or all R (so e.g ARA is allowed but RRR is not). Note that $X_i|Z_i = A$ has density $p(\cdot)$. This follows because

$$\begin{aligned}\mathbb{P}(X_i \leq x \,|\, A) = \mathbb{P}\left(X_i \leq x \,\middle|\, U_i \leq \tfrac{p(X_i)}{kr(X_i)}\right) &= \frac{\mathbb{P}\left(X_i \leq x\,,\, U_i \leq \frac{p(X_i)}{kr(X_i)}\right)}{\mathbb{P}\left(U_i \leq \frac{p(X_i)}{kr(X_i)}\right)} \\ &= \frac{\int_{-\infty}^{x}\left[\int_0^{p(x_i)/kr(x_i)} dy\right] r(x_i)\,dx_i}{\int_{-\infty}^{\infty}\left[\int_0^{p(x_i)/kr(x_i)} dy\right] r(x_i)\,dx_i} = \frac{\frac{1}{k}\int_{-\infty}^{x} p(x_i)\,dx_i}{\frac{1}{k}\int_{-\infty}^{\infty} p(x_i)\,dx_i} = \int_{-\infty}^{x} p(x_i)\,dx_i.\end{aligned}$$

(Note that this also gives a proof as to why you do not need to know the normalizing constant of $p(\cdot)$ if you only know the form of the density up to scaling - why?)

(a) Find the density of $X_i|Z_i = R$. **Hint:** Proceed in the same way as above.

(b) Consider the estimator

$$\delta := \frac{1}{N}\left(\sum_{i\,:\,Z_i=A} h(X_i) + \sum_{i\,:\,Z_i=R} h(X_i)\frac{(k-1)p(X_i)}{kr(X_i) - p(X_i)}\right).$$

Explain carefully how δ can be interpreted as an importance sampling estimate of θ.