Models

Justin Salmon[1] and George Lancaster[2]

[1] *wr18313*
[2] *qv18258*

October 25, 2018

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Fusce maximus nisi ligula. Morbi laoreet ex ligula, vitae lobortis purus mattis vel. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Donec ac metus ut turpis mollis placerat et nec enim. Duis tristique nibh maximus faucibus facilisis. Praesent in consequat leo. Maecenas condimentum ex rhoncus, elementum diam vel, malesuada ante.

1 The Prior

1.1 Theory

Question 1.1

A Gaussian likelihood encodes the inherent noise present in most real-world data. In other words, most probabilistic processes in nature tend to be noisy, and that noise tends to follow a Gaussian distribution. Hence this is generally a good place to start.

Noise in input data - assume noise is gaussian -> implies gaussian likelihood

Question 1.2

Choosing a spherical covariance matrix essentially means that we are assuming that the input noise is going to be equal in all directions, i.e. the distribution is equally likely to deviate from the mean in all directions. Again, this is a good place to start. Choosing a non-spherical covariance would imply that we know something in advance about the relationship between the input and output, which is not true in this case.

Question 2

TODO The covariance matrix would not be in terms of the identity matrix (for some reason). We would have non-zero values in the offset diagonals which correspond to the correlations between different things.

1.1.1 Linear Regression

Question 3

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \beta) = \prod_{i=0}^{N} \mathcal{N}(y_i|\mathbf{W}^T \phi(x_i), \beta^{-}1) \qquad (1)$$

Question 4

A distribution is conjugate to another if they both take the same algebraic form, i.e. are in the same probability distribution family. For example, Gaussians are conjugate to each other, and the conjugate to a Bernoulli distribution is a Beta distribution. Conjugates are used as a convenience used to avoid calculating the denominator in Baye's rule (the evidence) which can often be a tricky integral. If the prior and likelihood are conjugate, then their product will be proportional to the posterior.

Question 5

Euclidean distance from the mean

X appears only in the gaussian exponential part

Question 6

Derive posterior mean and covariance, start at conjugacy. Watch video https://www.youtube.com/watch?v=nrd4AnDLR3U

1.1.2 Non-parametric Regression

Question 7

Non-parametric: not defining parameters, just using what we know about the current data to classify new data points The simplest example of a. nonparementric model is a K-nearest neieghbour nmodel. The data are the parameters

Interpretability:

Question 8

This prior represents the space of all possible functions.

We think that smooth functions are more likely, but non-smooth functions e.q. sawtooth are non-zero but very unlikely

Question 9

All possible functions? Yes, but some are more likely than others

Question 10

$$p(\mathbf{Y}, \mathbf{X}, f, \theta) = p(Y|f)p(F|X, \theta)p(X)p(\theta) \tag{2}$$

TODO draw graphical model

Assumptions: - X and theta do not depend on anything - F depends on theta and X - Y depends on F

Question 11

The marginal in Eq. 2 connects the prior and the data because we now have a way to directly generate Y values given X and theta value without knowing the actual form of f.

The uncertainty "filters" through this ... We are uncertain about the function, so by marginalising it out the uncertainty gets "baked in" to Y.

The fact that θ is left on the LHS of the expression after marginalisation means that it is needed (along with x) to calculate Y. Y is dependent on theta.

Use the model diagram to help?

Marginalisation is the average of some distributions It is a weighted average of some functions, based on how likely each function is

1.2 Practical

1.2.1 Linear Regression

Question 12.1

2 Posterior

Question 15

Assumptions are what we assume to be true about something. They are a starting point, a way of being able to move forward from a point given no data about the problem. If we had all the data in the world, we would need no assumptions.

Use the assumptions to formulate a prior, use the prior to formulate your beliefs. Believe that the smooth lines are more likely. Prefer straight or flat lines.

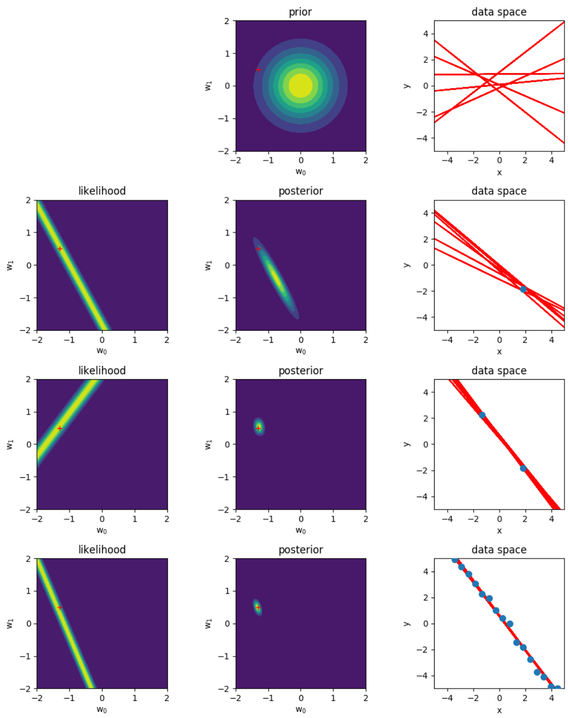

Figure 1: *TODO.*

Assumptions and preferences are very much similar. Preferences are the assumptions you want to use the most because it is easier exploiting conjugacy)

Beliefs inform assumptiosn (gaussian mean 0)

Question 16

We have assumed that the latent input variables are independent (since the covariance matrix is spherical) and we assume (or prefer) that they are centred around the origin (zero mean).

Question 17

3 Evidence