Machine Learning and Pattern Recognition - Week 4 - Bayesian Regression

Antonio León Villares

October 2022

Contents

Based on the online notes here.

1 Gaussians for Regression

- **Why is it a good idea to look at regression from a probabilistic perspective?**
  - data is often **noisy**
  - using **probabilistic models** allows us to gauge our uncertainty about the data
  - for instance, given some training data, there are a variety of functions which would be a good fit, whilst least squares only gives us one such possibility

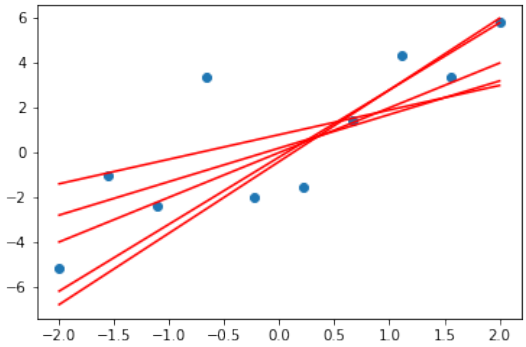


Figure 1: The above data was generated by $y = 2x + \mathcal{N}(\mu = 0, \sigma^2 = 4)$. Even the line $y = 2x$ doesn't fully explain the points. There is always uncertainty associated with the noisy data; least squares is just a "heuristic" to decide on a **possible** explanation.

- **How can we define a probabilistic model?**
  - let $\underline{x}$ be input, and y be the observation
  - lets assume that there are a set of weights $\underline{w}$ which define the function generating the data:

    $$f(\underline{x}; \underline{w})$$

    f could be a linear model with basis function, a NN, etc...
  - however, there is also noise associated with the actual observed output, defined by σ_y^2 (we assume this is known and applicable to each $\underline{x}$)
  - our **probabilistic model** evaluates how likely it is to see the observations y, given the data $\underline{x}$ and the weights $\underline{w}$:

    $$P(y \mid \underline{w}, \underline{x})$$
  - for example, a **Gaussian** model would be:

    $$P(y \mid \underline{w}, \underline{x}) = \mathcal{N}(y; \mu = f(\underline{x}; \underline{w}), \sigma^2 = \sigma_y^2)$$

- **How can we find the optimal weights $\underline{w}$?**
  - this is a probabilistic model, so we can use **negative log-likelihood**

– if $\underline{y}$ are all our observations, and X is our design matrix (where each observation is assumed to be independent):

$$\begin{aligned}
-\log(P(\underline{y} \mid X, \underline{w})) &= -\log\left(\prod_{n=1}^{N} P(y^{(n)} \mid \underline{x}^{(n)}, \underline{w})\right) \\
&= -\sum_{n=1}^{N} \log(P(y^{(n)} \mid \underline{x}^{(n)}, \underline{w})) \\
&= -\sum_{n=1}^{N} \log(\mathcal{N}(y^{(n)}; f(\underline{x}^{(n)}; \underline{w}), \sigma_y^2)) \\
&= -\sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(y^{(n)} - f(\underline{x}^{(n)}, \underline{w}))^2}{2\sigma_y^2}}\right) \\
&= -\sum_{n=1}^{N} \left(\log\left(\frac{1}{\sqrt{2\pi\sigma_y^2}}\right) - \frac{(y^{(n)} - f(\underline{x}^{(n)}, \underline{w}))^2}{2\sigma_y^2}\right) \\
&= -\sum_{n=1}^{N} \left(-\frac{1}{2}\log(2\pi\sigma_y^2) - \frac{1}{2\sigma_y^2}(y^{(n)} - f(\underline{x}^{(n)}, \underline{w}))^2\right) \\
&= \frac{N}{2}\log(2\pi\sigma_y^2) + \frac{1}{2\sigma_y^2}\sum_{n=1}^{N}(y^{(n)} - f(\underline{x}^{(n)}, \underline{w}))^2
\end{aligned}$$

– thus, if σ_y is known and constant, finding $\underline{w}$ is equivalent to fitting least squares (this can be thought of as a justification for applying it in linear regression)

- **What happens to the loss if the noise σ_y varies?**
  - this can occur if for example the measurement instrument has varying **precision** (i.e a thermometer might be more certain at lower temperatures)
  - then we would optimise:

$$\sum_{n=1}^{N}\left(\frac{1}{2}\log(2\pi(\sigma_y^{(n)})^2) + \frac{1}{2(\sigma_y^{(n)})^2}(y^{(n)} - f(\underline{x}^{(n)}, \underline{w}))^2\right)$$

  - in fact, the term:

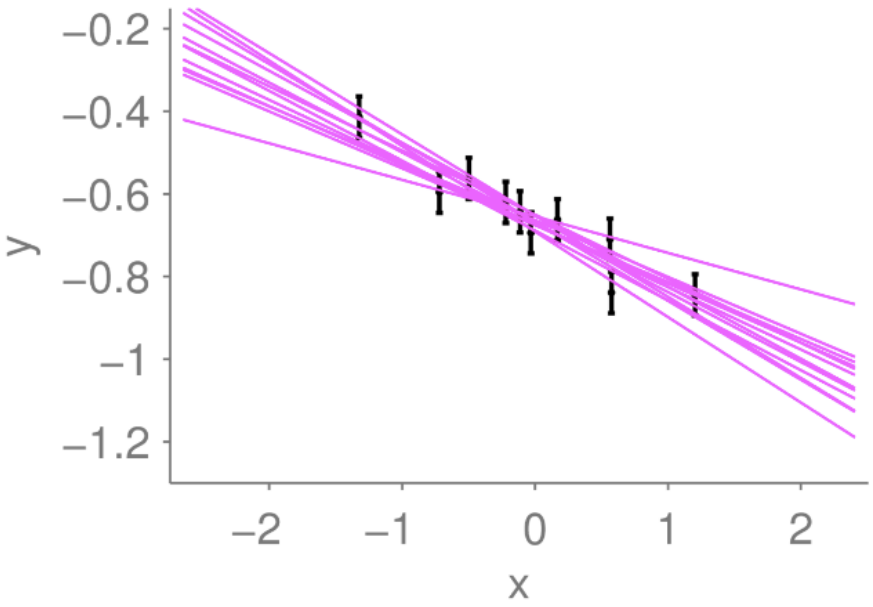
$$\frac{1}{(\sigma_y^{(n)})^2}$$

    is known as the **precision**
  - the above loss shows that observations with higher precision are weighted more than those with lower precision when computing the difference of squares

*It is important to note that this method is very similar to the standard linear regression: ultimately, we are optimising for a single set of weights, whilst a probabilistic model should consider **all** possible weights.*
However, we use this as an illustrative example of how probability can be used for regression models.
*The next sections go into **Bayesian Methods**, which are used to generate a distribution over models, which truly gauge the uncertainty in our observations derived from noise.*

2 Representing Uncertainty in Regression

- **How can we represent the noise in plots?**
  - we can use **error bars**, of width σ_y
  - a line would be good "fit" if it passes through $\approx 68\%$ of the error bars
  - assuming that noise is Gaussian, 68% of observations should lie within $\pm\sigma$ of the mean, so such a line would be a decent predictor
- **How can we use probability to represent our uncertainty about the regression task?**
  - say we have some noisy data, and some (possible) models explaining the data:



  - we can think of the model parameters $\underline{w}$ as our **beliefs**: they represent what we think can be a model
  - we can use **probability distributions** to represent **beliefs** (i.e we can think that each weight $\underline{w}$ is **sampled** from some distribution)
  - using **probability theory**, we can **update** these beliefs by using **new observations** (i.e if we observe more data, we can more accurately narrow down the distribution from which to sample $\underline{w}$, such that the resulting $\underline{w}$ better explain our new observations)

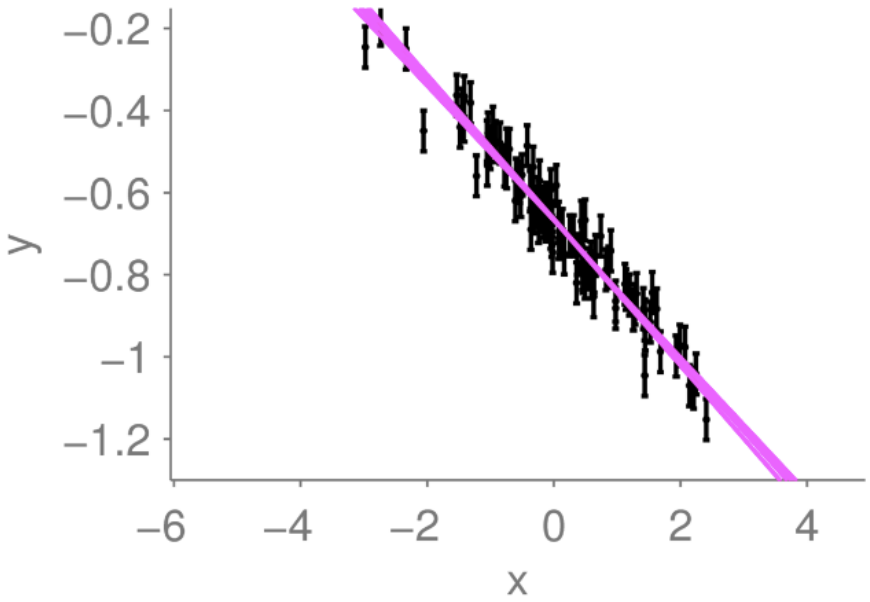
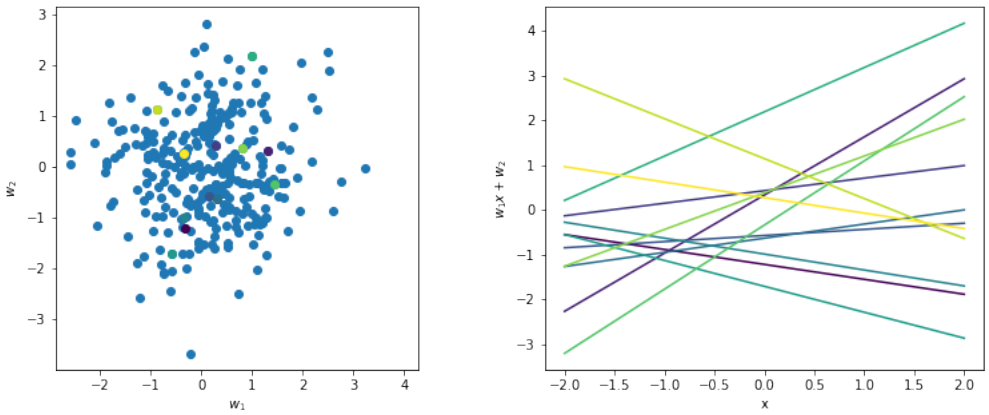


Figure 2: After adding more data points, our beliefs (the weights) can be updated. Notice that now the lines are more "focused" on the data, since we have less uncertainty: the distribution from which we draw $\underline{w}$ will have lower variance.

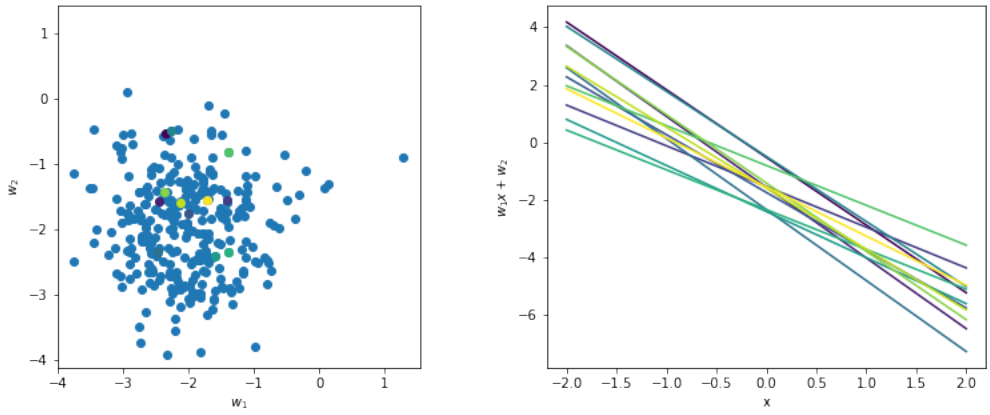
- **How can we generate models from a distribution?**
  - say we sample $\underline{w}$ from a normal distribution:

$$\underline{w} \sim \mathcal{N}(\underline{0}, \mathbb{I})$$

  - we can then use each sampled $\underline{w}$ to define a line $\underline{w}^T \underline{x}$:



  - if we then observe new data such as the one above (where the gradient is clearly negative), we can update our beliefs (i.e sampling distribution) to one where the mean is negative, and the variance is lower (since we have now observed data):



- **What is a prior belief?**
  - the **distribution** from which we assume that $\underline{w}$ is **sampled**
  - doesn't consider any data - just represents the models which are **plausible**
  - for example, if we sample from a Gaussian:

$$P(\underline{w}) = \mathcal{N}(\underline{w}; \underline{\mu} = \underline{0}, \Sigma = 0.4^2\mathbb{I})$$

    we obtain the following, plausible, models

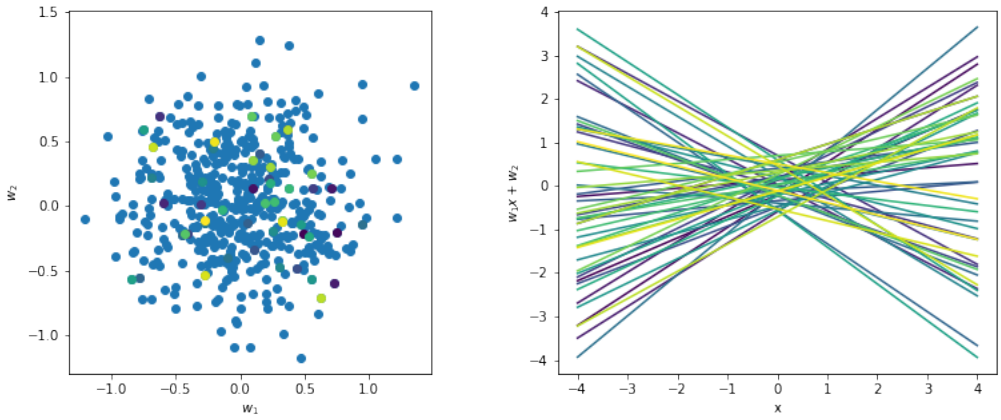


Figure 3: Generating models by sampling 500 $\underline{w}$ (plotted 50). For larger y-intercepts, we should change the variance associated to w_2.

- **How can we determine whether a prior leads to plausible distributions?**
  - if we haven't observed any data, it might be hard to define a **prior** - it should generate plausible models, but we don't even know what we are modelling

1. **Domain Knowledge**: from experience, we might know the general shape of data (i.e features tend to be positively correlated), and thus can select distributions which generate this sort of data (i.e a Gaussian with a mean vector with positive terms)

2. **Mathematical Convenience**: certain distributions make the math tractable. For instance, we might choose a gaussian with very high variance to estimate a **uniform** distribution, since gaussians are easier to work with.
3. **First Principles**: the **central limit theorem** states that, given enough data, distributions tend to look Gaussian, so a Gaussian is not a bad prior to pick!

- **What is a posterior belief?**
  - the **distribution** illustrating how our **beliefs** (weights) change after we actually observe data
  - if we have data:
    $$\mathcal{D} = \{\underline{x}^{(n)}, y^{(n)}\}$$
    the **posterior** is the distribution:
    $$P(\underline{w} \mid \mathcal{D})$$
    (that is, how likely we are to sample certain weights, now that we have seen the data which we want to fit)
- **How can the posterior distribution be computed?**
  - we use **Bayes' Rule**:
    $$P(\underline{w} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \underline{w})P(\underline{w})}{P(\mathcal{D})} \propto P(\mathcal{D} \mid \underline{w})P(\underline{w})$$
  - $P(\mathcal{D} \mid \underline{w})$ is the **likelihood**: how likely we are to observe $\mathcal{D}$ given a set of weights $\underline{w}$. It is **not** a distribution over the model parameters:
    $$\int P(\mathcal{D} \mid \underline{w}) d\underline{w} \neq 1$$
    but we can think of it as a function, which maps $\underline{w}$ to a probability of $\underline{w}$ generating $\mathcal{D}$
- **How can the posterior distribution be computed with the objective of regression?**
  - in regression, the **data** is just our observed values (the target):
    $$\mathcal{D} = \underline{y} = \{y^{(n)}\}$$
  - this data is actually **conditioned** on the inputs in the **design matrix** X (or Φ if we use basis functions), since we are assuming that X somehow generates $\underline{y}$
  - thus, we rewrite the posterior:
    $$P(\underline{w} \mid \mathcal{D}) = P(\underline{w} \mid \underline{y}, X) = \frac{P(\underline{y} \mid \underline{w}, X)P(\underline{w})}{P(\underline{y} \mid X)} \propto P(\underline{y} \mid \underline{w}, X)P(\underline{w})$$
  - hence, our **current beliefs** (**posterior**) are determined by our **original beliefs** about the possible models (**prior**), and how likely it is that a model generates the observed data (**likelihood**)

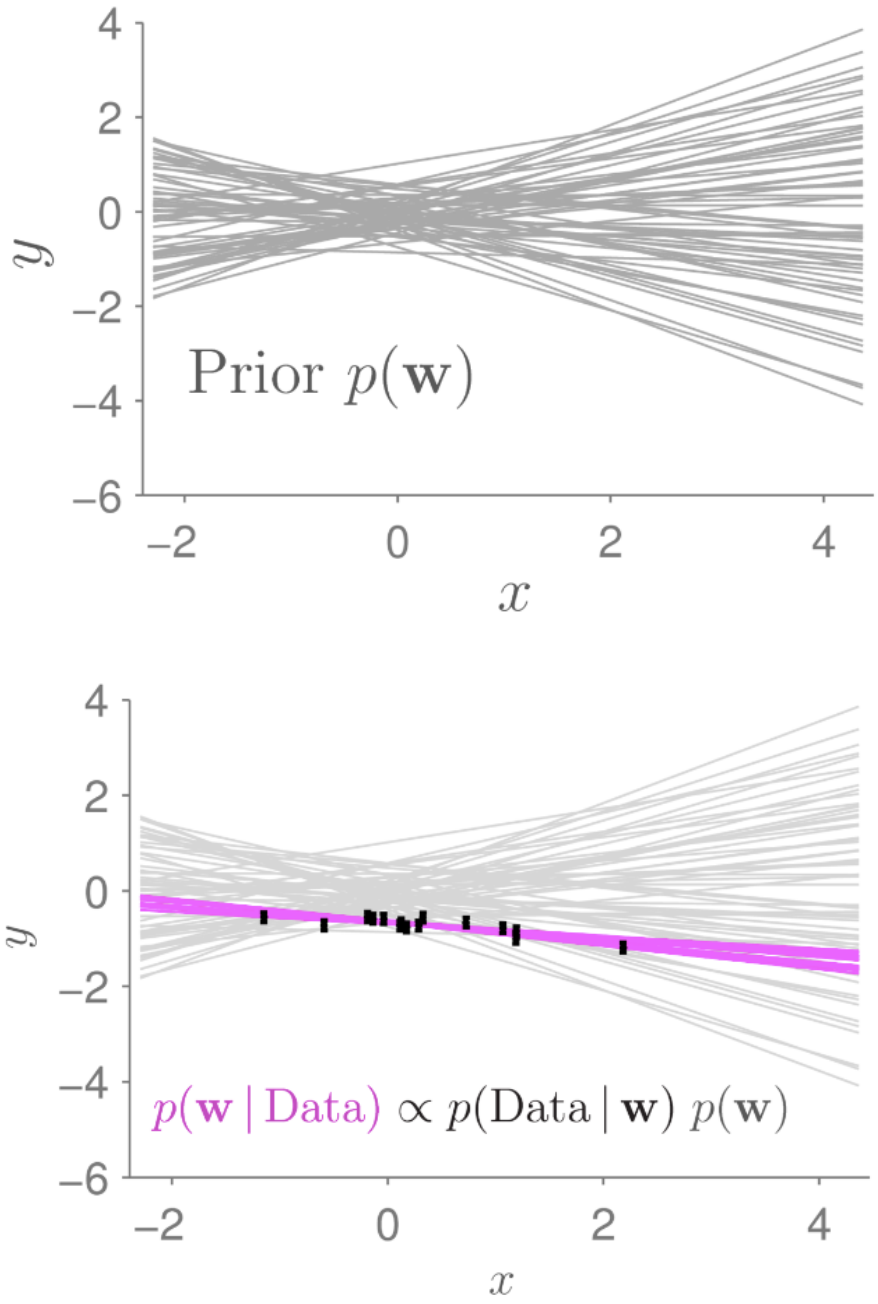


Figure 4: After observing data, sampling from the posterior distribution leads to models which more closely fit the data. We can interpret this as the variance (uncertainty) getting much smaller as we update our beliefs. Notice, away from the data, the models tend to "spread" indicating the uncertainty present when away from observed data.

- **What are conjugate priors?**
  - we have the formula:

$$posterior \propto likelihood \times prior$$

  - if the **posterior** and **prior** are in the **same** family of distributions (i.e both normal, both uniform, both gamma, etc...), then the **prior** is a **conjugate prior** to the **likelihood**

- for linear regression, if the likelihood and prior over the weights are **Gaussian**, then the **posterior** will be Gaussian (see below)

- **What is the closed-form distribution for the posterior?**
  - we have that:
    $$P(\underline{w} \mid \mathcal{D}) \propto P(\underline{w})P(\underline{y} \mid \underline{w}, X)$$
  - let:
    $$P(\underline{w}) = \mathcal{N}(\underline{w}; \underline{\mu} = \underline{w}_0, \Sigma = V_0)$$
    where the underscripts indicate that these are prior parameters (we haven't yet observed anything)
  - moreover:
    $$P(\underline{y} \mid \underline{w}, X) = \mathcal{N}(\underline{y}; \Phi\underline{w}, \sigma_y^2\mathbb{I})$$
    since conditioning on $X, \underline{w}$ is the same as using the model as our mean, and the noise as the variance (covariance) (here we use Φ, since the original data X might've been transformed by basis functions)
  - hence:
    $$\begin{aligned} P(\underline{w}; \mathcal{D}) &\propto \mathcal{N}(\underline{w}; \underline{\mu} = \underline{w}_0, \Sigma = V_0)\mathcal{N}(\underline{y}; \Phi\underline{w}, \sigma_y^2\mathbb{I}) \\ &\propto exp\left(-\frac{1}{2}(\underline{w} - \underline{w}_0)^T V_0^{-1}(\underline{w} - \underline{w}_0)\right) exp\left(-\frac{1}{2}(\underline{y} - \Phi\underline{w})^T(\sigma_y^2)\mathbb{I}(\underline{y} - \Phi\underline{w})\right) \\ &= exp\left(-\frac{1}{2}\left((\underline{w} - \underline{w}_0)^T V_0^{-1}(\underline{w} - \underline{w}_0) + (\underline{y} - \Phi\underline{w})^T(\sigma_y^2)\mathbb{I}(\underline{y} - \Phi\underline{w})\right)\right) \end{aligned}$$
  - if we define:
    $$P(\underline{w} \mid \mathcal{D}) = \mathcal{N}(\underline{w}; \underline{w}_N, V_n)$$
    then we get that:
    $$V_n = \sigma_y^2(\sigma_y^2 V_0^{-1} + \Phi^T\Phi)^{-1}$$
    $$\underline{w}_n = V_n V_0^{-1}\underline{w}_0 + \frac{1}{\sigma_y^2}V_n\Phi^T\underline{y}$$

3 Predictions with Bayesian Linear Regression

3.1 Motivation: A Simple Card Game

3.1.1 Problem Setup

We consider 3 cards, labelled 1,2,3:

- card 1 has 1 white and 1 black side
- card 2 has 2 black sides
- card 3 has 2 white sides

We shuffle the cards, turning them over randomly in the process. We pick a card, and upon placing it on the table, we see a black side. **What is the probability that the other side of the same card is white?**

Intuitively, we can say that the probability is $\frac{1}{3}$, since flipping a card and seeing a different colour is only possible if we picked card 1, and we will pick this card with a probability of $\frac{1}{3}$. However, this argument won't work for more complex problems, so we seek to "formalise" this sort of reasoning.

3.1.2 Conditioning on Cards

When solving inference problems, we should seek to first have a good model of the data.

The distribution of cards is:

$$P(C = c) = \frac{1}{3}, \qquad c \in \{1, 2, 3\}$$

Now, we define the probability of seeing a black face first, given we pick a given card $c \in \{1, 2, 3\}$:

$$P(F = B \mid C = c) = \begin{cases} \frac{1}{2}, & c = 1 \\ 1, & c = 2 \\ 0, & c = 3 \end{cases}$$

Using Baye's Rule can then tell us the probability of having picked a certain card, given that we saw a black face. Since card 1 is the only card with black and white faces, $P(C = 1 \mid F = B)$ defines the probability we are seeking. Thus:

$$P(C = c \mid F = B) = \frac{P(F = B \mid C = c)P(C = c)}{P(F = B)} = \begin{cases} \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}, & c = 1 \\ \frac{1 \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}, & c = 2 \\ 0, & c = 3 \end{cases}$$

so as expected $P(C = 1 \mid F = B) = \frac{1}{3}$.

Notice, the distribution is **unbalanced**: this makes sense, since card 2 is more likely to have a black face shown than card 1.

3.1.3 Formalising the Argument

We can further generalise this argument. In particular, the probability which we should've sought was:

$$P(F_2 = W \mid F_1 = B)$$

but this probability is impossible to find directly using Bayes' Theorem (we would find that $P(F_2 = W \mid F_1 = B) = P(F_1 = B \mid F_2 = W)$ which is rather unhelpful). We were able to solve the problem because we were only dealing with 3 cards and some simple rules.

If we want to find $P(F_2 = W \mid F_1 = B)$, we need to exploit the **product rule**:

$$P(X, Y) = P(X|Y)P(Y)$$

and the **sum rule**:

$$P(X) = \sum_{Y} P(X, Y)$$

Then, by the sum rule, we can introduce the card:

$$P(F_2 = W \mid F_1 = B) = \sum_{c \in C} P(F_2 = W, C = c \mid F_1 = B)$$

By the product rule, we can split this into a product, conditioned by the card type:

$$P(F_2 = W \mid F_1 = B) = \sum_{c \in C} P(F_2 = W \mid F_1 = B, C = c)P(C = c \mid F_1 = B)$$

These are all easily computable probabilities, and we will indeed get $\frac{1}{3}$!

In general, if we want to solve a prediction problem:

$$P(y \mid data)$$

we can do so by introducing a latent variable, and applying the product+sum rule combination:

$$P(y \mid data) = \sum_{z \in Z} P(y \mid z, data) P(z \mid data)$$

or for continuous RVs:

$$P(y \mid data) = \int P(y \mid z, data) P(z \mid data) dz$$

3.2 Linear Regression via Bayesian Models

- **What is the posterior predictive distribution?**
  - a **distribution** enabling us to predict an output y, given a new input $\underline{x}$, given some prior training data:
    $$\mathcal{D} = \{\underline{x}^{(n)}, y^{(n)}\}$$
  - we can compute $P(y \mid \underline{x}, \mathcal{D})$ by **conditioning** on the weights $\underline{w}$ of our model:
    $$P(y \mid \underline{x}, \mathcal{D}) = \int P(y, \underline{w} \mid \underline{x}, \mathcal{D}) d\underline{w} = \int P(y \mid \underline{x}, \underline{w}) P(\underline{w} \mid \mathcal{D}) d\underline{w}$$
  - notice, we drop $\mathcal{D}$ from the first term, since if we know $\underline{w}$, the training data doesn't influence the prediction
  - similarly, the second term isn't conditioned on $\underline{x}$, since we don't use the input $\underline{x}$ shouldn't tell us anything about the weights (unless we do transductive learning)
- **Can we compute the value of the integral?**
  - in general, this is hard; since we are using Gaussians, the integral actually has a closed form
  - notice, $P(y \mid \underline{x}, \underline{w})$ is our predictive distribution, given parameters:
    $$P(\underline{y} \mid \underline{x}, \underline{w}) = \mathcal{N}(y; \mu = \underline{w}^T \underline{x}, \sigma^2 = \sigma_y^2)$$
  - moreover, $P(\underline{w} \mid \mathcal{D})$ is our **posterior**:
    $$P(\underline{w} \mid \mathcal{D}) = \mathcal{N}(\underline{w}; \underline{\mu} = \underline{w}_N, \Sigma = V_N)$$
  - thus, the distribution to predict y given $\underline{x}$ is:
    $$P(y \mid \underline{x}, \mathcal{D}) = \int \mathcal{N}(y; \underline{w}^T \underline{x}, \sigma_y^2) \mathcal{N}(\underline{w}; \underline{\mu} = \underline{w}_N, \Sigma = V_N) d\underline{w}$$
  - whilst computable, this is tedious work
- **How can we compute $P(\underline{y} \mid \underline{x}, \mathcal{D})$ without explicitly computing the integral?**
  - we can exploit the fact that we expect a Gaussian output

- in particular, we can rewrite the **posterior predictive distribution** as:

$$y = \underline{x}^T \underline{w} + \nu, \qquad \nu \sim \mathcal{N}(0, \sigma_y^2)$$

- then, the parameters of the distribution will be given by:

$$\mathbb{E}[y] = \mathbb{E}[\underline{x}^T \underline{w}] + \mathbb{E}[\nu] = \mathbb{E}[\underline{x}^T \underline{w}] + 0 = \mathbb{E}[\underline{x}^T \underline{w}]$$

$$Var[y] = Var[\underline{x}^T \underline{w}] + Var[\nu] = Var[\underline{x}^T \underline{w}] + \sigma_y^2$$

(where we use the fact that $\underline{x}^T \underline{w}$ and ν are independent

- we thus compute:

$$\mathbb{E}[\underline{x}^T \underline{w}] = \underline{x}^T \mathbb{E}[\underline{w}] = \underline{x}^T \underline{w}_N$$

$$\begin{aligned} Var[\underline{x}^T \underline{w}] &= \mathbb{E}[(\underline{x}^T \underline{w} - \underline{x}^T \underline{w}_N)(\underline{x}^T \underline{w} - \underline{x}^T \underline{w}_N)] \\ &= \mathbb{E}[\underline{x}^T (\underline{w} - \underline{w}_N) \underline{x}^T (\underline{w} - \underline{w}_N)] \\ &= \mathbb{E}[\underline{x}^T (\underline{w} - \underline{w}_N)(\underline{w} - \underline{w}_N)^T \underline{x}], \qquad \textit{(by symmetry of inner product)} \\ &= \underline{x}^T \mathbb{E}[(\underline{w} - \underline{w}_N)(\underline{w} - \underline{w}_N)^T] \underline{x} \\ &= \underline{x}^T Cov[\underline{w}] \underline{x} \\ &= \underline{x}^T V_N \underline{x} \end{aligned}$$

- thus:

$$P(\underline{y} \mid \mathcal{D}, \underline{x}) = \mathcal{N}(y; \mu = \underline{x}^T \underline{w}_N, \sigma^2 = \underline{x}^T V_N \underline{x} + \sigma_y^2)$$

- notice, this implies that if we scale a datapoint $\underline{x}$ by $a\underline{x}$, the variance $\underline{x}^T V_N \underline{x} + \sigma_y^2$ will increase to:

$$a^2 \underline{x}^T V_N \underline{x} + \sigma_y^2$$

so the model will be less precise

3.3 Decision Making via Bayesian Models

- **How can we derive a single value from the posterior predictive distribution?**
  - notice, $P(y \mid \underline{x}, \mathcal{D})$ gives a **distribution**: a range of possible values for y
  - however, for **regression**, we typically want a **single** value
  - what this value is depends on the **loss** function $L(y, \hat{y})$, which indicates how bad our guess $\hat{y}$ will be
  - the **expected loss** given $\mathcal{D}$ gives us a **cost function**:

$$C = \mathbb{E}_{P(y \mid \underline{x}, \mathcal{D})}[L(y, \hat{y})] = \int L(y, \hat{y}) P(y \mid \underline{x}, \mathcal{D}) dy$$

  - by setting:

$$\frac{\partial C}{\partial \hat{y}} = 0$$

we can derive what our Bayesian model should output, to minimise the loss

- **What does a Bayesian Model output with a squared loss?**
  - let:

$$L(y, \hat{y}) = (y - \hat{y})^2$$

– then, we can differentiate inside the integral (since it doesn't depend on $\hat{y}$):

$$\begin{aligned}\frac{\partial C}{\partial \hat{y}} &= \int \frac{\partial}{\partial \hat{y}}(y-\hat{y})^2 P(y \mid \underline{x}, \mathcal{D}) dy \\ &= \int -2(y-\hat{y}) P(y \mid \underline{x}, \mathcal{D}) dy \\ &= -2\left(\int y P(y \mid \underline{x}, \mathcal{D}) dy - \int \hat{y} P(y \mid \underline{x}, \mathcal{D})\right) \\ &= -2(\mathbb{E}_{P(y \mid \underline{x}, \mathcal{D})}[y] - \hat{y})\end{aligned}$$

where we have used:

$$\int P(y \mid \underline{x}, \mathcal{D}) dy = 1$$

– in other words, if we have a Gaussian model, to minimise cost, we should predict:

$$\hat{y} = \mathbb{E}_{P(y \mid \underline{x}, \mathcal{D})}[y] = \underline{x}^T \underline{w}_N$$

It can be shown that this corresponds to using a L^2 regularised model, so using a Bayesian approach doesn't really change our linear regression. However, introducing this uncertainty is useful: an uncertain prediction can be used to manually inspect the data to understand where an issue might lie.

Imagine that a bakery makes an estimate of tomorrow's bread sales . If it were possible to make perfect predictions, so we knew that , then we would bake exactly loaves of bread. We'd sell all our bread, and send no customers away. Sadly we can't make perfect predictions. If we really don't like throwing bread away, then we could set smaller than the average possible value of . In that case, we would sell out nearly every day. Most bakers seem to attach a different loss to waste, and make decisions that regularly result in throwing bread away6.
As another example, a business-to-business supplier of non-perishable goods will pay warehouse fees to keep excess stock, if failing to fulfill orders will lose customers in the long term. For them, the cost of underestimating sales is much higher than the cost of overestimating.
The Bayesian approach separates modelling data from the application-specific loss function by keeping track of multiple possibilities for the model, and deferring making a decision until later. Multiple decisions, with different losses, can be made from the same model. An alternative and popular approach, "empirical risk minimization", fits a model function to the training data to directly optimize an application-specific loss function.

4 Tutorial

1. **We consider a probabilistic model for regression:**

$$P(y \mid \underline{x}, \underline{w}) = \mathcal{N}(y; f(\underline{x}; \underline{w}), \sigma_y^2)$$

In this question, we set $f(\underline{w}; \underline{w}) = \underline{w}^T \underline{x}$ and assume a diagonal prior covariance matrix:

$$P(\underline{w}) = \mathcal{N}(\underline{w}; \underline{w}_0, \sigma_w^2 \mathbb{I})$$

Recall, the posterior distribution is Normal:

$$P(\underline{w} \mid \mathcal{D}) = \mathcal{N}(\underline{w}; \underline{w}_N, V_N)$$

where:

$$V_N = \sigma_y^2(\sigma_y^2 V_0^{-1} + \Phi^T \Phi)^{-1}$$
$$\underline{w}_N = V_N V_0^{-1} \underline{w}_0 + \frac{1}{\sigma_y^2} V_N \Phi^T \underline{y}$$

(recall, $\underline{w}_0, V_0$ are the parameters for the prior distribution).
Using $\Phi = X$ and $V_0 = \sigma_w^2 \mathbb{I}$, we get that:

$$V_N = \sigma_y^2(\sigma_y^2 \frac{1}{\sigma_w^2} \mathbb{I} + X^T X)^{-1}$$
$$\underline{w}_N = V_N \frac{1}{\sigma_w^2} \mathbb{I} \underline{w}_0 + \frac{1}{\sigma_y^2} V_N X^T \underline{y}$$

What distribution, with what parameters is $P(\underline{w} \mid \mathcal{D})$ approaching when we let $\sigma_w^2 \to \infty$? Can you justify this intuitively?

- as $\sigma_w^2 \to \infty$, the posterior becomes a normal distribution with parameters:

$$V_N = \sigma_y^2 (X^T X)^{-1}$$
$$\underline{w}_N = (X^T X)^{-1} X^T \underline{y}$$

- notice, $(X^T X)^{-1} X^T$ is the Moore-Penrose Pseudo Inverse, so the mean of the posterior distribution will be the weights obtained by least squares fitting
- this makes intuitive sense: as $\sigma_w^2 \to \infty$, we are infinitely uncertain about the possible distribution of weights; hence, we just pick a mean which best fits to the data, without considering any prior information

2. N **noisy independent observations are made of the unknown saclar quantity m:**

$$x^{(n)} \sim \mathcal{N}(m, \sigma^2)$$

(a) **We don't give yout he raw data, but tell you the mean of the observations:**

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x^{(n)}$$

What is the $P(\bar{x} \mid m)$ (i.e what is the likelihood of M given $\bar{x}$).

Since $\bar{x}$ is a (scaled) sum of normally distributed RVs, $\bar{x}$ must itself be a normally distributed RV. Hence, determining the likelihood just requires finding the parameters of the distribution:

$$\mathbb{E}[\bar{x}] = \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} x^{(n)}\right] = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[x^{(n)}] = m$$

$$Var[\bar{x}] = Var\left[\frac{1}{N}\sum_{n=1}^{N} x^{(n)}\right] = \frac{1}{N^2}\sum_{n=1}^{N} Var[x^{(n)}] = \frac{\sigma^2}{N}$$

Thus:

$$P(\bar{x} \mid m) = \mathcal{N}\left(\bar{x}; m, \frac{\sigma^2}{N}\right)$$

(b) **A *sufficient statistic* is a summary of some data that contains all of the information about a parameter.**

i. **Show that $\bar{x}$ is a sufficient statistic of the observations for m, assuming we know the noise variance σ^2. That is, show that:**

$$P(m \mid \bar{x}) = P(m \mid \{x^{(n)}\})$$

We shall treat m as the only variable here, since we know both σ^2 and $\bar{x}$. We can then use proportionality to "ignore" any terms which don't include m.
We compute using Bayes Rule:

$$\begin{aligned}
P(m \mid \bar{x}) &\propto P(\bar{x} \mid m)P(m) \\
&\propto P(m)\exp\left(-\frac{1}{2}\frac{(\bar{x}-m)^2}{\frac{\sigma^2}{N}}\right) \\
&= P(m)\exp\left(-\frac{N}{2\sigma^2}(\bar{x}^2 - 2m\bar{x} + m^2)\right) \\
&\propto P(m)\exp\left(-\frac{Nm^2}{2\sigma^2} + \frac{mN\bar{x}}{\sigma^2}\right)
\end{aligned}$$

Similarly:

$$\begin{aligned}
P(m \mid \{x^{(n)}\}) &\propto P(\{x^{(n)}\} \mid m)P(m) \\
&\propto P(m)\prod_n P(x^{(n)} \mid m, \sigma^2) \\
&\propto P(m)\prod_n \exp\left(-\frac{1}{2}\frac{(x^{(n)}-m)^2}{\sigma^2}\right) \\
&= P(m)\exp\left(-\frac{1}{2\sigma^2}\sum_n (x^{(n)}-m)^2\right) \\
&= P(m)\exp\left(-\frac{1}{2\sigma^2}\sum_n ((x^{(n)})^2 - 2mx^{(n)} + m^2)\right) \\
&\propto P(m)\exp\left(-\frac{1}{2\sigma^2}\sum_n ((-2mx^{(n)} + m^2)\right) \\
&\propto P(m)\exp\left(-\frac{Nm^2}{2\sigma^2} + \frac{m\sum_n x^{(n)}}{\sigma^2}\right) \\
&= P(m)\exp\left(-\frac{Nm^2}{2\sigma^2} + \frac{mN\bar{x}}{\sigma^2}\right)
\end{aligned}$$

Hence, $\bar{x}$ is a sufficient statistic: our beliefs about the parameter m given $\bar{x}$ are identical to those given that we have the whole dataset

ii. **If we don't know the noise variance σ^2 or the mean, is $\bar{x}$ still a sufficient statistic?**

- intuitively, this isn't the case; the normal distribution is fully determined by its mean and variance
- if we don't know the variance, we won't know how spread out the data is (i.e it could all be sharply distributed about the mean, or it could be very noisy and spread out with the same mean)
- for instance, say all the samples are identically $\bar{x}$; then, we'd have 0 variance, and our distribution would be a delta distribution
- if otherwise there is high variance, but still mean $\bar{x}$, we'd eb uncertain about $\bar{x}$ as the mean
- $\bar{x}$ is still a good point-estimate, but it doesn't tell us the whole story about the distribution

3. **Recall, a *conjugate prior* for a likelihood function is a prior where the posterior is a distribution in the same family as the prior. For instance, a Gaussian prior on the mean of a Gaussian distribution is conjugate to Gaussian observations of that mean.**

   (a) **The inverse-gamma distribution is a distribution over positive numbers. It's often used to put a prior on the variance of a Gaussian distribution, because it's a conjugate prior.**
   The inverse-gamma distribution has a PDF:

$$P(z \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} \exp\left(-\frac{\beta}{z}\right), \qquad \alpha, \beta > 0$$

   Assume we obtain N observations:

$$x^{(n)} \sim \mathcal{N}(0, \sigma^2)$$

   where the variance is unknown. Say we place an inverse-gamma prior with parameters α, β on the variance. Show that the posterior of the variance, given the data, is also inverse-gamma, and find its parameters.

   Using Bayes' Rule (and using $v = \sigma^2$):

$$\begin{aligned} P(v \mid \{x^{(n)}\}) &\propto P(\{x^{(n)}\} \mid v) P(v) \\ &= P(\{x^{(n)}\} \mid 0, v) P(v \mid \alpha, \beta) \\ &= \left[\prod_n P(x^{(n)} \mid 0, v)\right] \frac{\beta^\alpha}{\Gamma(\alpha)} v^{-\alpha-1} \exp\left(-\frac{\beta}{v}\right) \\ &= \frac{1}{v^{N/2}} \exp\left(-\frac{1}{2v}\sum_n (x^{(n)})^2\right) \frac{\beta^\alpha}{\Gamma(\alpha)} v^{-\alpha-1} \exp\left(-\frac{\beta}{v}\right) \\ &\propto v^{-\left(\alpha+\frac{N}{2}\right)-1} \exp\left(-\frac{\beta + \frac{1}{2}\sum_n (x^{(n)})^2}{v}\right) \end{aligned}$$

   Hence, the posterior for v follows an inverse gamma distribution, with:

$$\alpha = \alpha + \frac{N}{2}$$

$$\beta = \beta + \frac{1}{2}\sum_n (x^{(n)})^2$$

   (b) **If a conjugate prior exists, then the data can be replaced with sufficient statistics. Can you explain why?**
   - since we have a conjugate prior, the posterior is parametrised by some distribution, which just depends on its parameters (in the case of a normal distribution, mean and variance)
   - the posterior is fully determined by these parameters, so they will be sufficient statistics - sampled data is no longer necessary