Overview

Recall module is roughly split into four parts:

1. **Random events**: counting, events, axioms of probability, Bayes, independence
2. **Random variables**: discrete RVs, mean and variance, correlation, conditional expectation
3. **Mid-term**
4. **Inequalities and laws of large numbers**: Markov, Chebyshev, Chernoff bounds, sample mean, weak law of large numbers, central limit theorem, confidence intervals, bootstrapping
5. **Statistical models**: continuous random variables, logistic regression, least squares
Overview

- Regression vs Classification
- Maximum Likelihood Estimation
- Bayesian Estimation
- MAP Estimation
Regression vs Classification

In classification we have:

- Vector $\vec{x}$ of $m$ observed features $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ e.g. blood pressure, age, cholestrol
- Label $Y$ we are trying to predict, a finite set of possible values e.g. heart condition
- Model: Assumed statistical relationship between features $\vec{x}$ and label $Y$

Alternatively $Y$ is a continuous valued random variable, so may be real-valued, and:

- Prediction is now usually referred to as regression (rather than classification).
- Quantity $Y$ is often referred to as the output or dependent variable (rather than the label)
Linear models are very popular for regression as easy to work with.

- Assume a linear relationship between observed feature vector $\vec{x}$ and dependent variable $Y$

  $$Y = \sum_{i=1}^{m} \Theta^{(i)} x^{(i)} + M$$

  where $\vec{\Theta}$ is a vector of unknown (random) parameters and $M$ is random “noise”.

- Vector $\vec{\Theta}$ is unknown and we want to estimate it.
Linear Models

• To estimate $\vec{\Theta}$ we need some **training data** i.e.
  - A set of observations consisting of pairs of values $(\vec{x}_1, Y_1)$, $(\vec{x}_2, Y_2)$, $\ldots$, $(\vec{x}_n, Y_n)$
  - We assume that $Y_1 = \sum_{i=1}^{m} \Theta^{(i)} x_1^{(i)} + M_1$ where $M_1$ is noise,
    $Y_2 = \sum_{i=1}^{m} \Theta^{(i)} x_2^{(i)} + M_2$ etc
  - Observe that $\vec{\Theta}$ is the same for every pair of observations but the noise $M_1, M_2$ etc varies.
• Plus the prior distributions of $\vec{\Theta}$ and $M$. For now we will assume:
  - $M$ is Gaussian with mean 0 and variance 1, $\Theta^{(i)}$ is Gaussian with mean 0 and variance $\lambda$ (with the value of $\lambda$ known)
  - which we write in shorthand as $M \sim N(0, 1), \Theta^{(i)} \sim N(0, \lambda)$ (recall $Y$ is a Normal random variable $Y \sim N(\mu, \sigma^2)$ when it has PDF:
  
  $f_Y(y) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$.)
Linear Models

Example: \( Y = x + M \), linear model with only one input so \( m = 1 \).

- Training data generated using Matlab code: 
  \[
  x = [0:0.1:10]; \\
  y = x + \text{randn}(1, \text{length}(x)); \text{plot}(x, y, '+')
  \]
- In this example \( \vec{\Theta} \) has just one element \((m = 1)\) with value \( \Theta^{(1)} = 1 \). In practice we don’t know \( \Theta^{(1)} \) in advance, and given the training data we want to estimate its value.
Linear Models

Another example: \( Y = x^{(1)} + x^{(2)} + M \), linear model with two inputs so \( m = 2 \).

- Training data generated using Matlab code:
  \[
  \begin{align*}
  [x1,x2] &= \text{meshgrid}([0:0.2:10],[0:0.2:10]); \quad y = x1 + x2 + \text{randn(size(x1))}; \\
  \text{plot3}(x1,x2,y,'+')
  \end{align*}
  \]
Linear Models

Another example: generalised linear model.

- Suppose have single input $x$ and output is:
  \[ Y = \Theta^{(1)} x + \Theta^{(2)} x^2 + \cdots + \Theta^{(m)} x^m + M, \quad M \sim N(0, 1), \quad \Theta^{(i)} \sim N(0, \lambda) \]

- Define feature vector $\vec{z}$ with $z^{(1)} = x$, $z^{(2)} = x^2$, $\cdots$, $z^{(m)} = x^m$.

- Using this vector the model is:
  \[ Y = \sum_{i=1}^{m} \Theta^{(i)} z^{(i)} + M \]

Although model is nonlinear in $x$ it is linear in $\vec{z}$. These new $\vec{z}$ can be computed given input $x$, so it’s known.
Linear Models

Example: \( Y = x^2 + M \).

- Matlab code: 
  \[
  x = [0:0.1:10];
  y = x.^2 + 10*randn(1,length(x));
  plot(x,y,'+'); plot(x.^2,y,'+');
  \]
Linear Models

There’s another (clearer ?) way to write linear model in terms of PDFs.

- Previously used $Y = \mathbf{\Theta} x + M$, $M \sim N(0, 1)$, $\Theta^{(i)} \sim N(0, \lambda)$.
- Given $\mathbf{\Theta} = \mathbf{\theta}$, then $Y - \sum_{i=1}^{m} \theta(i) x(i) = M \sim N(0, 1)$ i.e.

$$f_{Y|X, \mathbf{\Theta}}(y|x, \mathbf{\theta}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\left(y - \sum_{i=1}^{m} \theta(i) x(i)\right)^2 / 2\right)$$

- Note that we have to use PDF rather than PMF since $Y$ is a continuous RV.
- Model also assumes $\Theta^{(i)} \sim N(0, \lambda)$ i.e.

$$f_{\Theta^{(i)}}(\theta) \propto \exp(-\theta^2 / 2\lambda)$$

- $f_{Y|X, \mathbf{\Theta}}(y|x, \mathbf{\theta})$ and $f_{\Theta^{(i)}}(\theta^{(i)})$ fully describe the linear model.
Parameter Estimation

• Recall Bayes Rule for PDFs

\[ f_{\Theta|D}(\vec{\theta}|d) = \frac{f_{D|\Theta}(d|\vec{\theta})f_{\Theta}(\vec{\theta})}{f_D(d)} \]

\textit{posterior} \quad \textit{likelihood} \quad \textit{prior}

• **Likelihood.** PDF of seeing the data \( d \) given model with parameter \( \Theta = \vec{\theta} \)

• **Prior.** Before seeing any data what is our belief about the model i.e. what is PDF of parameter values \( \Theta \).

• **Posterior.** After seeing the data, what is our belief about PDF of parameter values \( \Theta \) now that we have seen the data.
Maximum Likelihood Estimation

Select the value $\tilde{\theta}$ which maximises likelihood $f_{D|\Theta}(d|\tilde{\theta})$.

- $Y = \sum_{i=1}^{m} \Theta^{(i)} x^{(i)} + M, M \sim N(0, 1), \Theta^{(i)} \sim N(0, \lambda)$
- Conditioned on $\tilde{\Theta} = \tilde{\theta}$ we have:

$$f_{D|\Theta}(d|\tilde{\theta}) \propto L(\theta) = \exp\left(-\sum_{j=1}^{n}(y_j - \sum_{i=1}^{m} \theta^{(i)} x_j^{(i)})^2 / 2\right)$$

dropping the normalising constant as it doesn’t matter here.

- Take log (giving the “log-likelihood”):

$$\log f_{D|\Theta}(d|\tilde{\theta}) \propto \log L(\theta) = -\frac{1}{2} \sum_{j=1}^{n}(y_j - \sum_{i=1}^{m} \theta^{(i)} x_j^{(i)})^2$$

- We want to select $\tilde{\theta}$ to maximise $\log L(\tilde{\theta})$ i.e. the minimise

$$\sum_{j=1}^{n}(y_j - \sum_{i=1}^{m} \theta^{(i)} x_j^{(i)})^2$$

- Called the “least squares” estimate, for obvious reasons.
Maximum Likelihood Estimation: Warm up example

We have one input/feature and one output.

- Suppose our training data consists of just two observations: (1, 3), (2, 1)
- The log-likelihood is
  \[-\frac{1}{2} \sum_{j=1}^{2} (y_j - \theta^{(1)} x_j^{(1)})^2 = -\frac{1}{2} (1 - 3\theta^{(1)})^2 - \frac{1}{2} (2 - 1\theta^{(1)})^2\]
- What value of \(\theta^{(1)}\) maximises \(-\frac{1}{2} (1 - 3\theta^{(1)})^2 - \frac{1}{2} (2 - 1\theta^{(1)})^2\)?
Maximum Likelihood Estimation

We can find the maximum in closed form. Let’s work through example with linear model and only one input ($m = 1$).

- Select $\theta$ to maximise $\log L(\theta) = -\frac{1}{2} \sum_{j=1}^{n} (y_j - \theta x_j)^2$
- Compute derivative with respect to $\theta$:

$$\frac{dL}{d\theta} = \sum_{j=1}^{n} (y_j - \theta x_j)x_j = \sum_{j=1}^{n} y_jx_j - \theta \sum_{j=1}^{n} x_j^2$$

- Set derivative equal to 0 and solve for $\theta$:

$$\sum_{j=1}^{n} y_jx_j - \theta \sum_{j=1}^{n} x_j^2 = 0$$

$$\Rightarrow \quad \theta = \frac{\sum_{j=1}^{n} y_jx_j}{\sum_{j=1}^{n} x_j^2}$$
Maximum Likelihood Estimation

Revisit example: $Y = x + M$, linear model with one input.
Maximum Likelihood Estimation

Some observations:

- We select $\theta$ to minimise $\sum_{j=1}^{n} (y_j - \theta x_j)^2$, so “least squares”
- This comes directly from the fact that we modelled the noise $M$ as Gaussian. Choosing a different noise model would change what we minimise
- We get a “closed form” solution.
- Can be easily extended to when input is a vector $\vec{x}$ i.e. when $m > 1$, although we won’t do it here.
Example: Advertising Data

- Data taken from An Introduction to Statistical Learning with Applications in R (http://www-bcf.usc.edu/~gareth/ISL/data.html)
- Data consists of the advertising budgets for three media (TV, radio and newspapers) and the overall sales in 200 different markets.
Example: Advertising Data

- Least square linear fit
- Residuals are the difference between the value predicted by the fit and the measured value.
  - Do the residuals look “random” or do they have some “structure”? Is our model satisfactory?
  - We can use the residuals to estimate a confidence interval for the prediction made by our linear fit.
- We could use bootstrapping to estimate our confidence in the fit itself.
Example: Advertising Data

- How is the impact of the advertising spend on TV and radio related, if at all?
- Perhaps a quadratic fit would be better? If so, what does that imply for how we allocate our advertising budget?
Bayesian Estimation

- Estimate the posterior \( f_{\Theta|D}(\theta|d) \), rather than likelihood \( f_{D|\theta}(d|\theta) \)
- A distribution rather than just a single value.

Example:
- \( Y = \sum_{i=1}^{m} \Theta^{(i)} x^{(i)} + M, \ M \sim N(0, 1), \ \Theta^{(i)} \sim N(0, \lambda) \).
- Bayes theorem: \( f_{\Theta|D}(\bar{\theta}|d) = \frac{f_{D|\theta}(d|\bar{\theta}) f_{\theta}(\bar{\theta})}{f_{D}(d)} \)
- We already know that likelihood \( f_{D|\Theta}(d|\bar{\theta}) \propto \exp(-\sum_{j=1}^{n} (Y_j - \sum_{i=1}^{m} \theta^{(i)} x_j^{(i)})^2 / 2) \).
- From our model we have prior \( f_{\Theta^{(i)}}(\theta) \propto \exp(-\theta^2 / 2\lambda) \).
- \( f_{D}(d) \) is a normalising constant (so area under PDF \( f_{\Theta|D}(\bar{\theta}|d) \) is 1).
MAP Estimation

- Maximum a posteriori (MAP) estimation\(^1\)
- Select \(\theta\) that maximises posterior \(f_{\theta|D}(\theta|d)\) (back to a single value rather than a distribution).

![Graph showing MAP estimate](image)

- Runs into trouble if distribution has \(> 1\) peak.

---

\(^1\)Challenge: what does this Latin mean?
MAP Estimation

Let's work through example with linear model and only one input ($m = 1$).

- Example: $Y = \Theta x + M$, $\Theta, M \sim N(0, 1)$.
- Likelihood $f_{D|\Theta}(d|\theta) \propto \exp(-\sum_{j=1}^{n}(y_j - \theta x_j)^2/2)$, prior $f_{\Theta}(\theta) \propto \exp(-\theta^2/2\lambda)$, $f_D(d)$ does not depend on $\theta$.
- Taking log, we want to select $\theta$ to maximise the log-posterior
  
  $\log f_{\Theta|D}(\theta|d) \propto -\sum_{j=1}^{n}(y_j - \theta x_j)^2 - \theta^2/\lambda$

- Differentiate this with respect to $\theta$: $\sum_{j=1}^{n} 2(y_j - \theta x_j)x_j - 2\theta/\lambda$.
- Set derivative equal to 0:

  $\sum_{j=1}^{n} 2(y_j - \theta x_j)x_j - 2\theta/\lambda = 0$

  $\Rightarrow \theta = \frac{\sum_{j=1}^{n} y_j x_j}{1/\lambda + \sum_{j=1}^{n} x_j^2}$

- This is called "ridge regression".
MAP Estimation

- MAP estimate: \( \theta = \frac{\sum_{j=1}^{n} y_j x_j}{1/\lambda + \sum_{j=1}^{n} x_j^2} \)

- MAP estimate depends on our choice of \( \lambda \). Remember that this value reflects our prior belief (before seeing any observations) of the distribution of parameter \( \Theta \), \( f_\Theta(\theta) \propto \exp(-\theta^2/2\lambda) \).

- When \( \lambda = 0 \), then we are saying that we are certain \( \Theta \) is 0 (\( \frac{\sum_{j=1}^{n} y_j x_j}{1/\lambda + \sum_{j=1}^{n} x_j^2} \rightarrow 0 \) as \( \lambda \rightarrow 0 \)).

- When \( \lambda \) is very large we are saying that we know very little about the value of \( \Theta \) prior to making the observations. MAP estimate is then close to the maximum likelihood estimate (\( \frac{\sum_{j=1}^{n} y_j x_j}{1/\lambda + \sum_{j=1}^{n} x_j^2} \approx \frac{\sum_{j=1}^{n} y_j x_j}{\sum_{j=1}^{n} x_j^2} \))
MAP vs Maximum Likelihood Estimation

Difference between MAP and ML really kicks in when we only have a small number of observations, yet still need to make a prediction. Our prior beliefs are then especially important.
MAP vs Maximum Likelihood Estimation

- But as number $N$ of observations grows, impact of prior on posterior tends to decline.
- MAP estimate: $\theta = \frac{\sum_{j=1}^{n} n y_j x_j}{1/\lambda + \sum_{j=1}^{n} x_j^2}$. As $n$ grows, $\sum_{j=1}^{n} x_j^2 \gg 1/\lambda$. 

![Graph showing comparison between MAP and ML estimates](image-url)

<table>
<thead>
<tr>
<th>Number of observations</th>
<th>MAP estimate $\lambda=1/100$</th>
<th>ML estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>40</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>60</td>
<td>0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>80</td>
<td>0.9</td>
<td>1.1</td>
</tr>
<tr>
<td>100</td>
<td>1.0</td>
<td>1.2</td>
</tr>
</tbody>
</table>
MAP vs Maximum Likelihood Estimation

- Remember two interpretations of probability, as frequency and as belief
- Prior as belief
- Important when need to make a decision with limited data

http://xkcd.com/1132/
To continue ...

- Have a bash at reading: