Introduction to Statistical Machine Learning

Cheng Soon Ong & Christian Walder

Machine Learning Research Group
Data61 | CSIRO
and
College of Engineering and Computer Science
The Australian National University

Canberra
February – June 2019

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")
Part IV

Linear Regression 2
Linear Regression

- Basis functions
- Maximum Likelihood with Gaussian Noise
- Regularisation
Training and Testing

Training Phase

- training data \( x \)
- training targets \( t \)
- model with adjustable parameter \( w \)

Test Phase

- test data \( x \)
- test target \( t \)
- model with fixed parameter \( w^* \)

Fix the most appropriate \( w^* \)
**Bayesian Regression**

- **Bayes Theorem**

  \[
  \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{normalisation}}
  \]

  \[
  p(w \mid t) = \frac{p(t \mid w) p(w)}{p(t)}
  \]

  where we left out the conditioning on \(x\) (always assumed), and \(\beta\), which is assumed to be constant.

- **Likelihood** for i.i.d. data (\(\beta\), inverse variance of noise)

  \[
  p(t \mid w) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid y(x_n, w), \beta^{-1})
  \]

  \[
  = \prod_{n=1}^{N} \mathcal{N}(t_n \mid w^T \phi(x_n), \beta^{-1})
  \]

  \[
  = \text{const} \times \exp\{-\beta \frac{1}{2} (t - \Phi w)^T (t - \Phi w)\}
  \]

  \[
  = \mathcal{N}(t \mid \Phi w, \beta^{-1} I)
  \]
How to choose a prior?

- Can we find a prior for the given likelihood which
  - makes sense for the problem at hand
  - allows us to find a posterior in a ‘nice’ form

An answer to the second question:

**Definition (Conjugate Prior)**

A class of prior probability distributions $p(w)$ is conjugate to a class of likelihood functions $p(x \mid w)$ if the resulting posterior distributions $p(w \mid x)$ are in the same family as $p(w)$. 
Examples of Conjugate Prior Distributions

**Table**: Discrete likelihood distributions

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>Conjugate Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernoulli</td>
<td>Beta</td>
</tr>
<tr>
<td>Binomial</td>
<td>Beta</td>
</tr>
<tr>
<td>Poisson</td>
<td>Gamma</td>
</tr>
<tr>
<td>Multinomial</td>
<td>Dirichlet</td>
</tr>
</tbody>
</table>

**Table**: Continuous likelihood distributions

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>Conjugate Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>Pareto</td>
</tr>
<tr>
<td>Exponential</td>
<td>Gamma</td>
</tr>
<tr>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Multivariate normal</td>
<td>Multivariate normal</td>
</tr>
</tbody>
</table>
Conjugate Prior to a Gaussian Distribution

- Example: The Gaussian family is conjugate to itself with respect to a Gaussian likelihood function: if the likelihood function is Gaussian, choosing a Gaussian prior will ensure that the posterior distribution is also Gaussian.

- Given a marginal distribution for $x$ and a conditional Gaussian distribution for $y$ given $x$ in the form

  $$p(x) = \mathcal{N}(x \mid \mu, \Lambda^{-1})$$
  $$p(y \mid x) = \mathcal{N}(y \mid Ax + b, L^{-1})$$

- we get

  $$p(y) = \mathcal{N}(y \mid A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$$
  $$p(x \mid y) = \mathcal{N}(x \mid \Sigma\{A^T L(y - b) + \Lambda \mu\}, \Sigma)$$

  where $$\Sigma = (\Lambda + A^T LA)^{-1}.$$
Bayesian Regression

- Choose a Gaussian prior with mean $m_0$ and covariance $S_0$
  \[ p(w) = \mathcal{N}(w | m_0, S_0) \]

- After having seen $N$ training data pairs $(x_n, t_n)$, the posterior for the given likelihood is now
  \[ p(w | t) = \mathcal{N}(w | m_N, S_N) \]

  where
  \[
  m_N = S_N(S_0^{-1}m_0 + \beta \Phi^T t)
  \]
  \[
  S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi
  \]

- The posterior is Gaussian, therefore mode = mean.
- The maximum posterior weight vector $w_{MAP} = m_N$.
- Assume infinitely broad prior $S_0 = \alpha^{-1}I$ with $\alpha \to 0$, the mean reduces to the maximum likelihood $w_{ML}$. 
If we have not yet seen any data point ($N = 0$), the posterior is equal to the prior.

Sequential arrival of data points: Each posterior distribution calculated after the arrival of a data point and target value, acts as the prior distribution for the subsequent data point.

Nicely fits a sequential learning framework.
Bayesian Regression

- Special simplified prior in the remainder, \( m_0 = 0 \) and \( S_0 = \alpha^{-1}I \),
  \[
p(x \mid \alpha) = \mathcal{N}(x \mid 0, \alpha^{-1}I)
  \]
- The parameters of the posterior distribution \( p(w \mid t) = \mathcal{N}(w \mid m_N, S_N) \) are now
  \[
  m_N = \beta S_N \Phi^T t
  \]
  \[
  S_N^{-1} = \alpha I + \beta \Phi^T \Phi
  \]
- For \( \alpha \to 0 \) we get
  \[
  m_N \to w_{ML} = (\Phi^T \Phi)^{-1} \Phi^T t
  \]
- Log of posterior is sum of log likelihood and log of prior
  \[
  \ln p(w \mid t) = -\frac{\beta}{2} (t - \Phi w)^T (t - \Phi w) - \frac{\alpha}{2} w^T w + \text{const}
  \]
Bayesian Regression

- Log of posterior is sum of log likelihood and log of prior

\[
\ln p(w | t) = - \beta \frac{1}{2} (t - \Phi w)^T (t - \Phi w) + \frac{\alpha}{2} w^T w + \text{const}
\]

- Maximising the posterior distribution with respect to \( w \) corresponds to minimising the sum-of-squares error function with the addition of a quadratic regularisation term \( \lambda = \alpha / \beta \).
**Sequential Update of the Posterior**

- Example of a linear basis function model
- Single input $x$, single output $t$
- Linear model $y(x, w) = w_0 + w_1 x$.
- Data creation
  1. Choose an $x_n$ from the uniform distribution $\mathcal{U}(x \mid -1, 1)$.
  2. Calculate $f(x_n, a) = a_0 + a_1 x_n$, where $a_0 = -0.3$, $a_1 = 0.5$.
  3. Add Gaussian noise with standard deviation $\sigma = 0.2$,
     \[ t_n = \mathcal{N}(x_n \mid f(x_n, a), 0.04) \]
- Set the precision of the uniform prior to $\alpha = 2.0$. 


Sequential Update of the Posterior
Sequential Update of the Posterior
Predictive Distribution

- In the training phase, data $x$ and targets $t$ are provided.
- In the test phase, a new data value $x$ is given and the corresponding target value $t$ is asked for.
- Bayesian approach: Find the probability of the test target $t$ given the test data $x$, the training data $x$ and the training targets $t$.

$$p(t \mid x, x, t)$$

- This is the Predictive Distribution.
How to calculate the Predictive Distribution?

- Introduce the model parameter \( w \) via the sum rule

\[
p(t \mid x, x, t) = \int p(t, w \mid x, x, t)dw
\]

\[
= \int p(t \mid w, x, x, t)p(w \mid x, x, t)dw
\]

- The test target \( t \) depends only on the test data \( x \) and the model parameter \( w \), but not on the training data and the training targets

\[
p(t \mid w, x, x, t) = p(t \mid w, x)
\]

- The model parameter \( w \) are learned with the training data \( x \) and the training targets \( t \) only

\[
p(w \mid x, x, t) = p(w \mid x, t)
\]

- Predictive Distribution

\[
p(t \mid x, x, t) = \int p(t \mid w, x)p(w \mid x, t)dw
\]
Proof of the Predictive Distribution

- How to prove the Predictive Distribution in the general form?

\[ p(t \mid x, x, t) = \int p(t \mid w, x, x, t)p(w \mid x, x, t)dw \]

- Convert each conditional probability on the right-hand-side into a joint probability.

\[
\int p(t \mid w, x, x, t)p(w \mid x, x, t)dw \\
= \int \frac{p(t, w, x, x, t)}{p(w, x, x, t)} \frac{p(w, x, x, t)}{p(x, x, t)}dw \\
= \int \frac{p(t, w, x, x, t)}{p(x, x, t)}dw \\
= \frac{p(t, x, x, t)}{p(x, x, t)} \\
= \frac{p(t, x, x, t)}{p(t \mid x, x, t)} \\
= p(t \mid x, x, t)
\]
Predictive Distribution with Simplified Prior

- Find the predictive distribution

\[ p(t \mid t, \alpha, \beta) = \int p(t \mid w, \beta) p(w \mid t, \alpha, \beta) \, dw \]

(remember: The conditioning on the input variables \( x \) is often suppressed to simplify the notation.)

- Now we know (neglecting as usual to notate conditioning on \( x \))

\[ p(t \mid w, \beta) = \mathcal{N}(t \mid w^T \phi(x), \beta^{-1}) \]

- and the posterior was

\[ p(w \mid t, \alpha, \beta) = \mathcal{N}(w \mid m_N, S_N) \]

where

\[ m_N = \beta S_N \Phi^T t \]

\[ S_N^{-1} = \alpha I + \beta \Phi^T \Phi \]
If we do the convolution of the two Gaussians, we get for the predictive distribution

$$p(t | x, t, \alpha, \beta) = \mathcal{N}(t | m_N^T \phi(x), \sigma_N^2(x))$$

where the variance $\sigma_N^2(x)$ is given by

$$\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x).$$
Example with artificial sinusoidal data from $\sin(2\pi x)$ (green) and added noise. Number of data points $N = 1$.

Mean of the predictive distribution (red) and regions of one standard deviation from mean (red shaded).
Predictive Distribution with Simplified Prior

Example with artificial sinusoidal data from $\sin(2\pi x)$ (green) and added noise. Number of data points $N = 2$.

Mean of the predictive distribution (red) and regions of one standard deviation from mean (red shaded).
Predictive Distribution with Simplified Prior

Example with artificial sinusoidal data from $\sin(2\pi x)$ (green) and added noise. Number of data points $N = 4$.

Mean of the predictive distribution (red) and regions of one standard deviation from mean (red shaded).
**Predictive Distribution with Simplified Prior**

Example with artificial sinusoidal data from $\sin(2\pi x)$ (green) and added noise. Number of data points $N = 25$.

Mean of the predictive distribution (red) and regions of one standard deviation from mean (red shaded).
Predictive Distribution with Simplified Prior

Plots of the function $y(x, w)$ using samples from the posterior distribution over $w$. Number of data points $N = 1$. 
Predictive Distribution with Simplified Prior

Plots of the function $y(x, w)$ using samples from the posterior distribution over $w$. Number of data points $N = 2$. 

```latex
Plots of the function $y(x, w)$ using samples from the posterior distribution over $w$. Number of data points $N = 2$. 
```

```
Plots of the function $y(x, w)$ using samples from the posterior distribution over $w$. Number of data points $N = 2$. 
```
Plots of the function $y(x, w)$ using samples from the posterior distribution over $w$. Number of data points $N = 4$. 
Plots of the function $y(x, w)$ using samples from the posterior distribution over $w$. Number of data points $N = 25$. 
Limitations of Linear Basis Function Models

- Basis function $\phi_j(x)$ are fixed before the training data set is observed.
- Curse of dimensionality: Number of basis function grows rapidly, often exponentially, with the dimensionality $D$.
- But typical data sets have two nice properties which can be exploited if the basis functions are not fixed:
  - Data lie close to a nonlinear manifold with intrinsic dimension much smaller than $D$. Need algorithms which place basis functions only where data are (e.g., radial basis function networks, support vector machines, relevance vector machines, neural networks).
  - Target variables may only depend on a few significant directions within the data manifold. Need algorithms which can exploit this property (Neural networks).
Curse of Dimensionality

- Linear Algebra allows us to operate in $n$-dimensional vector spaces using the intuition from our 3-dimensional world as a vector space. No surprises as long as $n$ is finite.

- If we add more structure to a vector space (e.g. inner product, metric), our intuition gained from the 3-dimensional world around us may be wrong.

- Example: Sphere of radius $r = 1$. What is the fraction of the volume of the sphere in a $D$-dimensional space which lies between radius $r = 1$ and $r = 1 - \epsilon$?

- Volume scales like $r^D$, therefore the formula for the volume of a sphere is $V_D(r) = K_D r^D$.

$$
\frac{V_D(1) - V_D(1 - \epsilon)}{V_D(1)} = 1 - (1 - \epsilon)^D
$$
Curse of Dimensionality

- Fraction of the volume of the sphere in a $D$-dimensional space which lies between radius $r = 1$ and $r = 1 - \epsilon$

$$\frac{V_D(1) - V_D(1 - \epsilon)}{V_D(1)} = 1 - (1 - \epsilon)^D$$

![Graph showing the volume fraction as a function of $\epsilon$ for different dimensions $D$.](image)
**Curse of Dimensionality**

- Probability density with respect to radius $r$ of a Gaussian distribution for various values of the dimensionality $D$.

![Probability density plots for different dimensions](image)

- $D = 1$
- $D = 2$
- $D = 20$
Curse of Dimensionality

- Probability density with respect to radius \( r \) of a Gaussian distribution for various values of the dimensionality \( D \).
- Example: \( D = 2 \); assume \( \mu = 0, \Sigma = I \)

\[
\mathcal{N}(x \mid 0, I) = \frac{1}{2\pi} \exp \left\{ -\frac{1}{2} x^T x \right\} = \frac{1}{2\pi} \exp \left\{ -\frac{1}{2} (x_1^2 + x_2^2) \right\}
\]

- Coordinate transformation

\[
x_1 = r \cos(\phi) \\
x_2 = r \sin(\phi)
\]

- Probability in the new coordinates

\[
p(r, \phi \mid 0, I) = \mathcal{N}(r(x), \phi(x) \mid 0, I) \mid J \mid
\]

where \( |J| = r \) is the determinant of the Jacobian for the given coordinate transformation.

\[
p(r, \phi \mid 0, I) = \frac{1}{2\pi} r \exp \left\{ -\frac{1}{2} r^2 \right\}
\]
Curse of Dimensionality

- Probability density with respect to radius $r$ of a Gaussian distribution for $D = 2$ (and $\mu = 0, \Sigma = I$)

$$p(r, \phi | 0, I) = \frac{1}{2\pi} r \exp \left\{ -\frac{1}{2} r^2 \right\}$$

- Integrate over all angles $\phi$

$$p(r | 0, I) = \int_0^{2\pi} \frac{1}{2\pi} r \exp \left\{ -\frac{1}{2} r^2 \right\} d\phi = r \exp \left\{ -\frac{1}{2} r^2 \right\}$$

![Graph showing probability density with respect to radius r for different dimensions D = 1, 2, 20]
Summary: Linear Regression

- Basis functions
- Maximum likelihood with Gaussian noise
- Regularisation
- Bayesian linear regression
- Conjugate prior
- Sequential update of the posterior
- Predictive distribution
- Curse of dimensionality