Statistical Machine Learning

Christian Walder

Machine Learning Research Group
CSIRO Data61

and

College of Engineering and Computer Science
The Australian National University

Canberra
Semester One, 2020.

(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
- Bias variance decomposition
Training and Testing: (Non-Bayesian) Point Estimate

**Training Phase**
- **Training data** $x$
- **Training targets** $t$
- Model with adjustable parameter $w$
- Fix the most appropriate $w^*$

**Test Phase**
- **Test data** $x$
- **Test target** $t$
- Model with fixed parameter $w^*$
Bayesian Regression

Bayes Theorem

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

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

I.i.d. regression likelihood for additive Gaussian noise is

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

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

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

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

- The choice of prior affords an intuitive control over our inductive bias
- All inference schemes have such biases, and often arise more opaquely than the prior in Bayes’ rule.
- 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 (mean parameter)</td>
</tr>
<tr>
<td>Multivariate normal</td>
<td>Multivariate normal (mean parameter)</td>
</tr>
</tbody>
</table>
Conjugate Prior to a Gaussian Distribution

- Example: If the likelihood function is Gaussian, choosing a Gaussian prior for the mean 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

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

- we get

\[
p(y) = \mathcal{N}(y | A\mu + b, L^{-1} + A\Lambda^{-1}A^\top)
\]
\[
p(x | y) = \mathcal{N}(x | \Sigma \{A^\top L(y - b) + \Lambda \mu\}, \Sigma)
\]

where \( \Sigma = (\Lambda + A^\top LA)^{-1} \).

Note that the covariance \( \Sigma \) does not involve \( y \).
Conjugate Prior to a Gaussian Distribution (intuition)

Given

\[ p(x) = \mathcal{N}(x \mid \mu, \Lambda^{-1}) \]

\[ p(y \mid x) = \mathcal{N}(y \mid Ax + b, L^{-1}) \iff y = Ax + b + \mathcal{N}(0, L^{-1}) \]

We have \( \mathbb{E}[y] = \mathbb{E}[Ax + b] = A\mu + b \) and by the easily proven Bienaymé formula for the variance of the sum of uncorrelated variables,

\[ \text{cov}[y] = \underbrace{\text{cov}[Ax + b]}_{=\mathbb{E}[Ax(Ax)^\top] = \mathbb{E}[xx^\top]A^\top = A\Lambda^{-1}A^\top} + \underbrace{\text{cov}[\mathcal{N}(0, L^{-1})]}_{=L^{-1}} \]

So \( y \) is Gaussian with

\[ p(y) = \mathcal{N}(y \mid A\mu + b, L^{-1} + A\Lambda^{-1}A^\top) \]

Then letting \( \Sigma = (\Lambda + A^\top LA)^{-1} \) and

\[ p(x \mid y) = \mathcal{N}(x \mid \Sigma \{A^\top L(y - b) + \Lambda\mu\}, \Sigma) \]

\[ \iff x = \Sigma \{A^\top L(y - b) + \Lambda\mu\} + \mathcal{N}(0, \Sigma) \]

yields the correct moments for \( x \), since

\[ \mathbb{E}[x] = \mathbb{E}[\Sigma \{A^\top L(y - b) + \Lambda\mu\}] = \Sigma \{A^\top L(A\mu + b - b) + \Lambda\mu\} \]

\[ = \Sigma \{A^\top LA\mu + \Lambda\mu\} = (\Lambda + A^\top LA)^{-1} \{A^\top LA + \Lambda\} \mu = \mu, \]

and it is similar (but tedious; don’t do it) to recover \( \text{cov}[x] = \Lambda \).
Bayesian Regression

- Choose a Gaussian prior with mean $\mathbf{m}_0$ and covariance $S_0$
  \[ p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_0, S_0) \]
- Same likelihood as before (here written in vector form):
  \[ p(\mathbf{t} \mid \mathbf{w}, \beta) = \mathcal{N}(\mathbf{t} \mid \Phi^\top \mathbf{w}, \beta^{-1} I) \]
- Given $N$ data pairs $(\mathbf{x}_n, t_n)$, the posterior is
  \[ p(\mathbf{w} \mid \mathbf{t}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_N, S_N) \]
  where
  \[ \mathbf{m}_N = S_N(S_0^{-1} \mathbf{m}_0 + \beta \Phi^\top \mathbf{t}) \]
  \[ S_N^{-1} = S_0^{-1} + \beta \Phi^\top \Phi \]
  (derive this with the identities on the previous slides)
Bayesian Regression: Zero Mean, Isotropic Prior

- For simplicity we proceed with $m_0 = 0$ and $S_0 = \alpha^{-1}I$, so
  
  $$p(w | \alpha) = \mathcal{N}(w | 0, \alpha^{-1}I)$$

- The posterior becomes $p(w | t) = \mathcal{N}(w | m_N, S_N)$ with
  
  $$m_N = \beta S_N \Phi^\top t$$
  
  $$S_N^{-1} = \alpha I + \beta \Phi^\top \Phi$$

- For $\alpha \ll \beta$ we get
  
  $$m_N \to w_{ML} = (\Phi^\top \Phi)^{-1} \Phi^\top t$$

- Log of posterior is sum of log likelihood and log of prior
  
  $$\ln p(w | t) = -\frac{\beta}{2} (t - \Phi w)^\top (t - \Phi w) - \frac{\alpha}{2} w^\top w + \text{const}$$
Bayesian Regression

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

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

- The maximum a posteriori estimator

\[
w_{\text{m.a.p.}} = \arg \max_w p(w \mid t)
\]

This corresponds to minimising the sum-of-squares error function with quadratic regularisation coefficient \( \lambda = \alpha / \beta \).

- The posterior is Gaussian so mode = mean: \( w_{\text{m.a.p.}} = m_N \).

- For \( \alpha \ll \beta \) the we recover unregularised least squares (equivalently m.a.p. approaches maximum likelihood), for example in case of
  - an infinitely broad prior with \( \alpha \rightarrow 0 \)
  - an infinitely precise likelihood with \( \beta \rightarrow \infty \)
Bayesian Inference in General: Sequential Update of Belief

- If we have not yet seen any data point \((N = 0)\), the posterior is equal to the prior.

- Sequential arrival of data points: the posterior given some observed data acts as the prior for the future data.
- Nicely fits a sequential learning framework.
Sequential Update of the Posterior

- Example of a linear basis function model
- Single input $x$, single output $t$
- Linear model $y(x, w) = w_0 + w_1x$.
- True data distribution sampling procedure:
  1. Choose an $x_n$ from the uniform distribution $\mathcal{U}(x | -1, +1)$.
  2. Calculate $f(x_n, a) = a_0 + a_1x_n$, where $a_0 = -0.3$, $a_1 = 0.5$.
  3. Add Gaussian noise with standard deviation $\sigma = 0.2$,

$$t_n = \mathcal{N}(x_n | f(x_n, a), 0.04)$$

- Set the precision of the uniform prior to $\alpha = 2.0$. 
Sequential Update of the Posterior

- **likelihood**
- **prior/posterior**
- **data space**

Limitations of Linear Basis Function Models
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 $t$

$$p(t | x, x, t)$$

- This is the Predictive Distribution (c.f. the posterior distribution, which is over the parameters).
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

The predictive distribution is

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

because

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

\[ = \int \frac{p(t, w, x, x, t)}{p(x, x, t)} dw \]

\[ = \frac{p(t, x, x, t)}{p(x, x, t)} \]

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

or simply

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

\[ = p(t \mid x, x, t). \]
**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: conditioning on \(x\) is often suppressed to simplify the notation.)

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

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

- and the posterior was

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

where

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

\[
S_N^{-1} = \alpha I + \beta \Phi^\top \Phi
\]
If we do the integral (it turns out to be the convolution of the two Gaussians), we get for the predictive distribution

\[ p(t \mid x, t, \alpha, \beta) = \mathcal{N}(t \mid m_N^\top \phi(x), \sigma_N^2(x)) \]

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

\[ \sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^\top S_N \phi(x). \]

This is more easily shown using a similar approach to the earlier “intuition” slide and again with the Bienaymé formula, now using

\[ t = w^\top \phi(x) + \mathcal{N}(0, \beta^{-1}). \]

However this is a linear-Gaussian specific trick and in general we need to integrate out the parameters.
Predictive Distribution with Simplified Prior

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).
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).
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).
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 $N = 2$. 

![Plots of the function $y(x, w)$ using samples from the posterior distribution over $w$. Number of data points $N = 2$.](image)
Plots of the function $y(x, \mathbf{w})$ using samples from the posterior distribution over $\mathbf{w}$. Number of data points $N = 4$. 

![Plots showing the function $y(x, \mathbf{w})$](image-url)
Predictive Distribution with Simplified Prior

Plots of the function $y(x, w)$ using samples from the posterior distribution over $w$. Number of data points $N = 25$. 
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., kernel methods / Gaussian processes).
  - Target variables may only depend on a few significant directions within the data manifold. Need algorithms which can exploit this property (e.g., linear methods or shallow 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 $V_D(r) = K_D r^D$.

$$\frac{V_D(1) - V_D(1 - \epsilon)}{V_D(1)} = 1 - (1 - \epsilon)^D$$
Curve 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 for different dimensions](image)
Curse of Dimensionality

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

![Graph showing probability density for different dimensions](image)
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 \( \mu = 0, \Sigma = I \)

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

- Coordinate transformation

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

- Probability in the new coordinates

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

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 $\mu = 0, \Sigma = I$)

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

- Integrate over all angles $\phi$

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

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