LINEAR REGRESSION
Credits

- Some of these slides were sourced and/or modified from:
  - Christopher Bishop, Microsoft UK
Relevant Problems from Murphy

- 7.4, 7.6, 7.7, 7.9
- Please do 7.9 at least. We will discuss the solution in class.
Linear Regression Topics

- What is linear regression?
- Example: polynomial curve fitting
- Other basis families
- Solving linear regression problems
- Regularized regression
- Multiple linear regression
- Bayesian linear regression
What is Linear Regression?

- In classification, we seek to identify the *categorical* class $C_k$ associated with a given input vector $x$.
- In regression, we seek to identify (or *estimate*) a *continuous* variable $y$ associated with a given input vector $x$.
- $y$ is called the *dependent variable*.
- $x$ is called the *independent variable*.
- If $y$ is a vector, we call this multiple regression.
- We will focus on the case where $y$ is a scalar.
- Notation:
  - $y$ will denote the continuous model of the dependent variable
  - $t$ will denote discrete noisy observations of the dependent variable (sometimes called the *target variable*).
In regression we assume that $y$ is a function of $x$. The exact nature of this function is governed by an unknown parameter vector $w$:

$$y = y(x, w)$$

The regression is linear if $y$ is linear in $w$. In other words, we can express $y$ as

$$y = w^t \phi(x)$$

where

$\phi(x)$ is some (potentially nonlinear) function of $x$. 

Linear Basis Function Models

- Generally

\[ y(x, w) = \sum_{j=0}^{M-1} w_j \phi_j(x) = w^T \phi(x) \]

- where \( \phi_j(x) \) are known as basis functions.

- Typically, \( \Phi_0(x) = 1 \), so that \( w_0 \) acts as a bias.

- In the simplest case, we use linear basis functions: \( \Phi_d(x) = x_d \).
Linear Regression Topics

- What is linear regression?
- **Example:** polynomial curve fitting
- Other basis families
- Solving linear regression problems
- Regularized regression
- Multiple linear regression
- Bayesian linear regression
Example: Polynomial Bases

- Polynomial basis functions:
  \[ \phi_j(x) = x^j. \]

- These are global:
  - A small change in \( x \) affects all basis functions.
  - A small change in a basis function affects \( y \) for all \( x \).
Example: Polynomial Curve Fitting

\[ y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j \]
Sum-of-Squares Error Function

\[ E(w) = \frac{1}{2} \sum_{n=1}^{N} \left( y(x_n, w) - t_n \right)^2 \]
1st Order Polynomial
3rd Order Polynomial

\[ M = 3 \]
9th Order Polynomial

\[ M = 9 \]
Regularization

- Penalize large coefficient values

\[
\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2
\]
Regularization

9\textsuperscript{th} Order Polynomial

\[ \ln \lambda = -18 \]
Regularization

9th Order Polynomial

\[ \ln \lambda = 0 \]
Regularization

9\textsuperscript{th} Order Polynomial

![Graph showing the behavior of a 9th order polynomial with training and test error vs. log lambda (\ln \lambda).]
Probabilistic View of Curve Fitting

- Why least squares?

- Model noise (deviation of data from model) as Gaussian i.i.d.

\[ p(t|x_0, w, \beta) = \mathcal{N}(t|y(x_0, w), \beta^{-1}) \]

where \( \beta \triangleq \frac{1}{\sigma^2} \) is the precision of the noise.
Maximum Likelihood

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

- We determine \( w_{ML} \) by minimizing the squared error \( E(w) \).

\[
\ln p(t|x, w, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \left\{ y(x_n, w) - t_n \right\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) \underbrace{\beta E(w)}_{\text{\beta E(w)}}
\]

- Thus least-squares regression reflects an assumption that the noise is i.i.d. Gaussian.
Maximum Likelihood

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

- We determine \( w_{ML} \) by minimizing the squared error \( E(w) \).

\[
\ln p(t|x, w, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) \underbrace{\beta E(w)}
\]

- Now given \( w_{ML} \), we can estimate the variance of the noise:

\[
\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, w_{ML}) - t_n\}^2
\]
Predictive Distribution

\[ p(t|x, \mathbf{w}_{\text{ML}}, \beta_{\text{ML}}) = \mathcal{N}(t|y(x, \mathbf{w}_{\text{ML}}), \beta_{\text{ML}}^{-1}) \]
MAP: A Step towards Bayes

- Prior knowledge about probable values of $w$ can be incorporated into the regression:

$$p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}w^T w\right\}$$

- Now the posterior over $w$ is proportional to the product of the likelihood times the prior:

$$p(w|x, t, \alpha, \beta) \propto p(t|x, w, \beta)p(w|\alpha)$$

- The result is to introduce a new quadratic term in $w$ into the error function to be minimized:

$$\beta \tilde{E}(w) = \frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2 + \frac{\alpha}{2}w^T w$$

- Thus regularized (ridge) regression reflects a 0-mean isotropic Gaussian prior on the weights.
Linear Regression Topics

- What is linear regression?
- Example: polynomial curve fitting
- Other basis families
- Solving linear regression problems
- Regularized regression
- Multiple linear regression
- Bayesian linear regression
Gaussian Bases

- Gaussian basis functions:
  \[ \phi_j(x) = \exp \left\{ -\frac{(x - \mu_j)^2}{2s^2} \right\} \]

- These are local:
  - a small change in \( x \) affects only nearby basis functions.
  - a small change in a basis function affects \( y \) only for nearby \( x \).
  - \( \mu_j \) and \( s \) control location and scale (width).

Think of these as interpolation functions.
Linear Regression Topics

- What is linear regression?
- Example: polynomial curve fitting
- Other basis families
- **Solving linear regression problems**
- Regularized regression
- Multiple linear regression
- Bayesian linear regression
Assume observations from a deterministic function with added Gaussian noise:

\[ t = y(x, w) + \epsilon \quad \text{where} \quad p(\epsilon|\beta) = \mathcal{N}(\epsilon|0, \beta^{-1}) \]

which is the same as saying,

\[ p(t|x, w, \beta) = \mathcal{N}(t|y(x, w), \beta^{-1}). \]

where

\[ y(x, w) = \sum_{j=0}^{M-1} w_j \phi_j(x) = w^T \phi(x) \]
Maximum Likelihood and Linear Least Squares

- Given observed inputs, \( X = \{x_1, \ldots, x_N\} \), and targets, \( t = [t_1, \ldots, t_N]^T \) we obtain the likelihood function

\[
p(t|X, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n | w^T \phi(x_n), \beta^{-1}).
\]
Taking the logarithm, we get

\[
\ln p(t|w, \beta) = \sum_{n=1}^{N} \ln \mathcal{N}(t_n|w^T \phi(x_n), \beta^{-1})
\]

\[
= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(w)
\]

where

\[
E_D(w) = \frac{1}{2} \sum_{n=1}^{N} \left\{ t_n - w^T \phi(x_n) \right\}^2
\]

is the sum-of-squares error.
Maximum Likelihood and Least Squares

- Computing the gradient and setting it to zero yields

\[ \nabla_w \ln p(t|w, \beta) = \beta \sum_{n=1}^{N} \left\{ t_n - w^T \phi(x_n) \right\} \phi(x_n)^T = 0. \]

- Solving for \( w \), we get

\[ w_{ML} = \left( \Phi^T \Phi \right)^{-1} \Phi^T t \]

- where

\[ \Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \cdots & \phi_{M-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & \cdots & \phi_{M-1}(x_N) \end{pmatrix}. \]
Linear Regression Topics

- What is linear regression?
- Example: polynomial curve fitting
- Other basis families
- Solving linear regression problems
- Regularized regression
- Multiple linear regression
- Bayesian linear regression
Regularized Least Squares

- Consider the error function:

\[ E_D(w) + \lambda E_W(w) \]

Data term + Regularization term

- With the sum-of-squares error function and a quadratic regularizer, we get

\[
\frac{1}{2} \sum_{n=1}^{N} \{t_n - w^T \phi(x_n)\}^2 + \frac{\lambda}{2} w^T w
\]

- which is minimized by

\[
w = \left(\lambda I + \Phi^T \Phi\right)^{-1} \Phi^T t.
\]

Thus the name ‘ridge regression’
Application: Colour Restoration
Application: Colour Restoration

Original Image

Red and Blue Channels Only

Predicted Image

Remove Green

Restore Green
Regularized Least Squares

- A more general regularizer:

\[
\frac{1}{2} \sum_{n=1}^{N} \{ t_n - w^T \phi(x_n) \}^2 + \frac{\lambda}{2} \sum_{j=1}^{M} |w_j|^q
\]

![Graphs showing different values of q](image)

\( q = 0.5 \quad q = 1 \quad q = 2 \quad q = 4 \)

Lasso Quadratic

(Least absolute shrinkage and selection operator)
Lasso generates sparse solutions.

**Iso-contours of data term** $E_D(w)$

**Iso-contour of regularization term** $E_W(w)$

Quadratic  

Lasso
Solving Regularized Systems

- Quadratic regularization has the advantage that the solution is closed form.
- Non-quadratic regularizers generally do not have closed form solutions.
- Lasso can be framed as minimizing a quadratic error with linear constraints, and thus represents a convex optimization problem that can be solved by quadratic programming or other convex optimization methods.
- We will discuss quadratic programming when we cover SVMs.
Linear Regression Topics

- What is linear regression?
- Example: polynomial curve fitting
- Other basis families
- Solving linear regression problems
- Regularized regression
- Multiple linear regression
- Bayesian linear regression
Multiple Outputs

- Analogous to the single output case we have:

\[
p(t|x, W, \beta) = \mathcal{N}(t|y(W, x), \beta^{-1}I)
= \mathcal{N}(t|W^T\phi(x), \beta^{-1}I).
\]

- Given observed inputs \( X = \{x_1, \ldots, x_N\} \), and targets \( T = [t_1, \ldots, t_N]^T \)
  we obtain the log likelihood function

\[
\ln p(T|X, W, \beta) = \sum_{n=1}^{N} \ln \mathcal{N}(t_n|W^T\phi(x_n), \beta^{-1}I)
= \frac{NK}{2} \ln \left( \frac{\beta}{2\pi} \right) - \frac{\beta}{2} \sum_{n=1}^{N} \|t_n - W^T\phi(x_n)\|^2.
\]
Multiple Outputs

- Maximizing with respect to $W$, we obtain

$$W_{ML} = \left( \Phi^T \Phi \right)^{-1} \Phi^T T.$$  

- If we consider a single target variable, $t_k$, we see that

$$w_k = \left( \Phi^T \Phi \right)^{-1} \Phi^T t_k = \Phi^\dagger t_k.$$  

- where $t_k = [t_{1k}, \ldots, t_{Nk}]^T$, which is identical with the single output case.
Some Useful MATLAB Functions

- **polyfit**
  - Least-squares fit of a polynomial of specified order to given data

- **regress**
  - More general function that computes linear weights for least-squares fit
Linear Regression Topics

- What is linear regression?
- Example: polynomial curve fitting
- Other basis families
- Solving linear regression problems
- Regularized regression
- Multiple linear regression
- Bayesian linear regression
Bayesian Linear Regression

Rev. Thomas Bayes, 1702 - 1761
Bayesian Linear Regression

- In least-squares, we determine the weights $\mathbf{w}$ that minimize the least squared error between the model and the training data.
- This can result in overlearning!
- Overlearning can be reduced by adding a regularizing term to the error function being minimized.
- Under specific conditions this is equivalent to a Bayesian approach, where we specify a prior distribution over the weight vector.
Bayesian Linear Regression

- Define a conjugate prior over $w$:
  
  $$p(w) = \mathcal{N}(w|m_0, S_0).$$

- Combining this with the likelihood function and matching terms, we obtain
  
  $$p(w|t) = \mathcal{N}(w|m_N, S_N)$$

- where

  $$m_N = S_N \left( S_0^{-1} m_0 + \beta \Phi^T t \right)$$

  $$S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi.$$
Bayesian Linear Regression

- A common choice for the prior is
  \[ p(w) = \mathcal{N}(w|0, \alpha^{-1}I) \]

- for which
  \[
  m_N = \beta S_N \Phi^T t \\
  S_N^{-1} = \alpha I + \beta \Phi^T \Phi.
  \]

- Thus \( m_N \) represents the ridge regression solution with \( \lambda = \alpha / \beta \)

- Next we consider an example ...
Bayesian Linear Regression

Example: fitting a straight line

0 data points observed
Bayesian Linear Regression

1 data point observed

Likelihood for \((x_1,t_1)\)

Posterior

Data Space
Bayesian Linear Regression

2 data points observed

Likelihood for \((x_2, t_2)\)

Posterior

Data Space
Bayesian Linear Regression

20 data points observed

Likelihood for \((x_{20}, t_{20})\)

Posterior

Data Space
Bayesian Prediction

- In least-squares, or regularized least-squares, we determine specific weights $w$ that allow us to predict a specific value $y(x,w)$ for every observed input $x$.

- However, our estimate of the weight vector $w$ will never be perfect! This will introduce error into our prediction.

- In Bayesian prediction, we model the posterior distribution over our predictions, taking into account our uncertainty in model parameters.
Predictive Distribution

- Predict $t$ for new values of $x$ by integrating over $w$:

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta)p(\mathbf{w}|\mathbf{t}, \alpha, \beta) \, d\mathbf{w}$$

$$= \mathcal{N}(t|m_N^T \phi(x), \sigma^2_N(x))$$

- where

$$\sigma^2_N(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x).$$
Example: Sinusoidal data, 9 Gaussian basis functions, 1 data point

Notice how much bigger our uncertainty is relative to the ML method!!
Example: Sinusoidal data, 9 Gaussian basis functions, 2 data points

\[
E[t \mid t, \alpha, \beta] \quad p(t \mid t, \alpha, \beta)
\]
Example: Sinusoidal data, 9 Gaussian basis functions, 4 data points

$$E\left[ t \mid t, \alpha, \beta \right] \quad p(t \mid t, \alpha, \beta)$$

Samples of $$y(x, w)$$
Example: Sinusoidal data, 9 Gaussian basis functions, 25 data points

\[ E\left[t \mid t, \alpha, \beta\right] \quad p\left(t \mid t, \alpha, \beta\right) \]

Samples of \( y(x, \mathbf{w}) \)
Linear Regression Topics

- What is linear regression?
- Example: polynomial curve fitting
- Other basis families
- Solving linear regression problems
- Regularized regression
- Multiple linear regression
- Bayesian linear regression