LINEAR MODELS
REGRESSION

I don't trust linear regressions when it's harder to guess the direction of the correlation from the scatter plot than to find new constellations on it.
Agenda

• The Regression Problem
  • Classification vs. Regression

• Linear Regression
  • The Least Squares Criterion and Solution
    • Residual Analysis

• Examples
  • Polynomial fitting
  • Cosine fitting

• Regularized Least Squares
  • Lasso vs. Ridge Regression
THE REGRESSION PROBLEM
## Classification vs. Regression

<table>
<thead>
<tr>
<th></th>
<th>Classification</th>
<th>Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong></td>
<td>Feature vector $X$</td>
<td>Feature vector $X$</td>
</tr>
<tr>
<td><strong>Output</strong></td>
<td>Label/category (discrete)</td>
<td>Value (continuous)</td>
</tr>
<tr>
<td><strong>MLE</strong></td>
<td>$\hat{\theta} = \arg\max_{\theta} \log p(\mathcal{D}</td>
<td>\theta)$</td>
</tr>
<tr>
<td><strong>Appropriate Models</strong></td>
<td>Decision Trees, SVM, neural network…</td>
<td>Decision Trees, SVM, neural network…</td>
</tr>
</tbody>
</table>

### MLE Formulations

- **Classification:** $\hat{\theta} = \arg\max_{\theta} \log p(\mathcal{D}|\theta)$
- **Regression:** $\hat{\theta} = \arg\max_{\theta} \log p(\mathcal{D}|\theta)$
Example – House Price Model

• Assume we want to know what the price of a house would be for a given size?
  • Given house of size x what is the price y=f(x) of the house?

<table>
<thead>
<tr>
<th>Square Feet (x)</th>
<th>House Price in $1000s (y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1400</td>
<td>245</td>
</tr>
<tr>
<td>1600</td>
<td>312</td>
</tr>
<tr>
<td>1700</td>
<td>279</td>
</tr>
<tr>
<td>1875</td>
<td>308</td>
</tr>
<tr>
<td>1100</td>
<td>199</td>
</tr>
<tr>
<td>1550</td>
<td>219</td>
</tr>
<tr>
<td>2350</td>
<td>405</td>
</tr>
<tr>
<td>2450</td>
<td>324</td>
</tr>
<tr>
<td>1425</td>
<td>319</td>
</tr>
<tr>
<td>1700</td>
<td>255</td>
</tr>
</tbody>
</table>
Graphical Representation

- The House Price scatter plot

- Which function $y=f(x)$ to choose?

- Assuming a *linear* connection between dependent and independent variables limits the search space
LINEAR REGRESSION
Maximum Likelihood Estimation

• The most common way to estimate parameters of a statistical model is by calculating the MLE:

$$\hat{\theta} = \arg \max_{\theta} p(y|x, \theta)$$

• Under the i.i.d. assumption, we derive the negative log-likelihood:

$$NLL(\theta) = -\log p(\mathcal{D}|\theta) = - \sum_{i=1}^{n} \log p(y_i|x_i, \theta)$$
Maximum Likelihood Estimation

- If we assume that
  \[ P(y|x, \theta) = \mathcal{N}(y|w^T x, \sigma^2) \]
- We get:
  \[
  NLL(\theta) = - \sum_{i=1}^{n} \log \left[ \left( \frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp \left( - \frac{1}{2\sigma^2} (y_i - w^T x_i)^2 \right) \right] = \frac{1}{2\sigma^2} SSE(w) - \frac{N}{2} \log(2\pi\sigma^2)
  \]
- Maximizing the log likelihood is equivalent to minimizing MSE:
  \[
  MSE(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2
  \]
A Linear Model for the House Prices

- One independent variable (house size) which “explains” the dependent variable (house price)

\[ y = \beta_0 + \beta_1 x + \varepsilon \]
Graphical Representation

$y = \beta_0 + \beta_1 x + \varepsilon$

Observed Value of $y$ for $x_i$

Predicted Value of $y$ for $x_i$

Intercept = $\beta_0$

Random Error for this $x$ value

Slope = $\beta_1$

$\varepsilon_i$
Estimated Regression Model

• The sample regression line provides an estimate of the population regression line

\[ \hat{y} = w_0 + w_1 x \]

The individual random error terms \( e_i \) have a mean of zero.
Linear Regression Assumptions

• The underlying relationship between the $x$ variable and the $y$ variable is linear

• Error values $\varepsilon$ are statistically independent

• The probability distribution of the errors is normal and independent of $x$
  With mean 0 and an equal but unknown variance for all values of $x$
House Price Model

Scatter Plot and Regression Line

Intercept = 98.248

\[
\text{house price} = 98.24833 + 0.10977 \text{ (square feet)}
\]

Slope = 0.10977
Residual Analysis

• Purposes
  • Examine for linearity assumption
  • Examine for constant variance for all levels of x
  • Evaluate normal distribution assumption

• Graphical Analysis of Residuals
  • Can plot residuals vs. x
  • Can create histogram of residuals to check for normality
Residual Analysis for Linearity

Not Linear

Linear
Residual Analysis for Constant Variance

Non-constant variance

Constant variance
Closed-form Solution

- The parameters $\mathbf{w}$ are obtained by minimizing the sum of the squared residuals

$$E(\mathbf{w}) = \| \mathbf{X} \mathbf{w} - \mathbf{y} \|^2 = \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i - y_i)^2$$
Closed-form Solution

- The parameters $\mathbf{w}$ are obtained by minimizing the sum of the squared residuals

$$ E(\mathbf{w}) = \| \mathbf{X} \mathbf{w} - \mathbf{y} \|^2 $$

$$ \nabla_{\mathbf{w}} E = 2\mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{y}) $$

- Setting the derivative to zero yields the necessary condition for minimum

$$ \mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y} $$

$$ \mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} $$

- The matrix $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the *Pseudo Inverse of $\mathbf{X}$*

*One may also solve using iterative methods (i.e. gradient descent)*
Basis function expansion

- Linear regression can be made to model non-linear relationships by replacing \( x \) with some non-linear function of the inputs \( \phi(x) \)

\[
y = w_0 + w\phi(x) + \epsilon
\]

- The model is still linear in the parameters \( w \) so it is still called linear regression
EXAMPLES
Polynomial fitting

• Given $L$ measurements $\{x_k, y_k\}_{k=1}^L$ originating from $n^{th}$ degree polynomial with additive white Gaussian noise $y = h(x) + \epsilon$

Where: $h(x) = \sum_{i=0}^{n} w_i x^i$ and $\epsilon \sim N(0, \sigma^2)$ i.i.d

Goal: Find the coefficients $\{w_i\}_{i=1}^{n}$ to best fit the polynomial $h(x)$
Polynomial fitting – cont.

• For each measurements \((x_k, y_k)\) we will fit the following model

\[ y_k = w_0 x_k^0 + w_1 x_k^1 + w_2 x_k^2 + \cdots + w_n x_k^n \quad \forall k = 1, 2, \ldots, L \]
The loss function

- The optimization goal is to minimize the mean squared error

\[
f(a) = \sum_{k=1}^{L} (y_k - (w_0 x_k^0 + w_1 x_k^1 + w_2 x_k^2 + \cdots + w_n x_k^n))^2
\]

Denoting \( X = \begin{bmatrix} x_1^0 & \cdots & x_1^n \\ \vdots & \ddots & \vdots \\ x_L^0 & \cdots & x_L^n \end{bmatrix} \), \( w = \begin{bmatrix} w_0 \\ \vdots \\ w_n \end{bmatrix} \), and \( y = \begin{bmatrix} y_1 \\ \vdots \\ y_L \end{bmatrix} \),

We can write the loss function in matrix form

- \( f(w) = (Xw - y)^T (Xw - y) = \|Xw - y\|_2^2 \)

- The solution (as before):
  \[
  w^* = (X^T X)^{-1} X^T y
  \]
Cosine fitting

- Given \( L \) measurements \( \{x_k, y_k\}_{k=1}^L \) originating from the sinusoidal signal \( h(x) \) with additive white Gaussian noise \( y = h(x) + n \)

Where: \( h(x) = A \cos(2\pi f_0 x_k + \phi) \) and \( n \sim N(0, \sigma^2) \ i.i.d \)

The frequency \( f_0 \) is given.

Goal: Find the amplitude \( A \) and the phase \( \phi \)
Model parameters

\[ h(x) = A \cos(2\pi f_0 x_k + \phi) \] and \( n \sim N(0, \sigma^2) \) i.i.d

We want to find the amplitude \( A \) and the phase \( \phi \)
The loss function

- The optimization goal is to minimize the mean squared error

\[ f(A, \phi) = \sum_{k=1}^{L} (y_k - A \cos(2\pi f_0 x_k + \phi))^2 \]

- This problem is not linear.
- What can we do in order to use linear model?
Converting to linear least squares

- We will use the following trigonometric identity:
  \[ \cos(\alpha + \beta) = \cos(\alpha) \cos(\beta) - \sin(\alpha) \sin(\beta) \]
  \[ A \cos(2\pi f_0 x_k + \phi) = A \cos(\phi) \cos(2\pi f_0 x_k) - A \sin(\phi) \sin(2\pi f_0 x_k) \]

- Next, we will use the following transformation from polar coordinates:
  - \( w_1 = A \cos(\phi) \)
  - \( w_2 = A \sin(\phi) \)
  - And get:
  - \( w_1 \cos(2\pi f_0 x_k) - w_2 \sin(2\pi f_0 x_k) \)
The loss function

• The optimization goal is to minimize the mean squared error

\[
f(A, \phi) = \sum_{k=1}^{L} \left( y_k - A \cos(2\pi f_0 x_k + \phi) \right)^2
\]

• This problem is not linear.

• With the suggested change of variables we get:

\[
f(w_1, w_2) = \sum_{k=1}^{L} \left( y_k - (w_1 \cos(2\pi f_0 x_k) - w_2 \sin(2\pi f_0 x_k)) \right)^2
\]

• This is a linear least squares problem!
The loss function

- The optimization goal is to minimize the mean squared error

\[ f(w_1, w_2) = \sum_{k=1}^{L} (y_k - (w_1 \cos(2\pi f_0 x_k) - w_2 \sin(2\pi f_0 x_k)))^2 \]

Denoting \( X = \begin{bmatrix} \cos(2\pi f_0 x_1) & -\sin(2\pi f_0 x_1) \\ \vdots & \vdots \\ \cos(2\pi f_0 x_L) & -\sin(2\pi f_0 x_L) \end{bmatrix} \)

\( w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_L \end{bmatrix} \)

We can write the loss function in a matrix form

- \( f(w) = (Xw - y)^T(Xw - y) = \|Xw - y\|_2^2 \)
The loss function

The loss function in a matrix form

\[ f(a) = (Xw - y)^T (Xw - y) = \|Xw - y\|^2 \]

And the solution is:

\[ w^* = (X^T X)^{-1} X^T y \]

In order to get \( A^* \) and \( \phi^* \) we will use the inverse transformation

\[ A = \sqrt{w_1^2 + w_2^2} \]

\[ \phi = \tan^{-1} \left( \frac{w_2}{w_1} \right) \]
REGULARIZATION
Bias–variance tradeoff

- Given a true (but unknown) function $F(x)$ with noise, we seek to estimate it based on $n$ samples from a set $\mathcal{D}$. Denoting the regression function $g(x; \mathcal{D})$.
- The error of our model is given by:
  $$MSE = E_\mathcal{D} \left[ (g(x; \mathcal{D}) - f(x))^2 \right]$$
- The error can be decomposed into two terms:
  $$E_\mathcal{D} \left[ (y - f(x))^2 \right] = \left( E_\mathcal{D} [g(x; \mathcal{D}) - F(x)] \right)^2 + E_\mathcal{D} \left[ (g(x; \mathcal{D}) - E_\mathcal{D} [g(x; \mathcal{D})])^2 \right]$$
  - $bias^2$
  - $variance$
Bias–variance tradeoff

- **Bias** - $E_D[g(x; D) - F(x)]$ - the difference between the expected value and the true (but generally unknown value).
  - Low bias = on average, we accurately estimate $F$ from $D$

- **Variance** - $E_D[(g(x; D) - E_D[g(x; D)])^2]$ - The portion of the error that is due to variation in data selection and finiteness of the dataset
  - Low Variance – the estimate of $F$ does not change much as the training set varies
Ridge regression

The loss function in a matrix form

\[ Loss = (Xw - y)^T (Xw - y) = \|Xw - y\|^2 \]

After adding the regularization term we get:

\[ Loss = \|Xw - y\|^2 + \lambda \|w\|^2 = (Xw - y)^T (Xw - y) + \lambda w^T w \]

The gradient of the loss with respect to \( w \) is:

\[ 2(X^T Xw - X^T y) + 2\lambda w \]

And the solution is:

\[ w^* = (X^T X + \lambda)^{-1} X^T y \]

The Shrinking Effect

When \( \lambda \to 0 \) we resort back to the (unconstrained) Least Squares solution

As \( \lambda \) increased, the num. of degrees of freedom of \((\lambda I + X^T X)^{-1} X^T\) will be reduced
LASSO regression

- \(\text{Ridge} = \|Xw - y\|_2^2 + \lambda \|w\|_2^2\)
- \(\text{LASSO} = \|Xw - y\|_2^2 + \lambda \|w\|_1\)

- The optimal weights are:
  - For ridge
    \(w^{ridge} = \text{argmin} \|Xw - y\|_2^2\)
    \(s.t. \quad \|w\|_2^2 < t\)
  - For LASSO
    \(w^{Lasso} = \text{argmin} \|Xw - y\|_2^2\)
    \(s.t. \quad \|w\|_1 < t\)
Other Regularization Terms

- With a more general regularizer, we have

$$\frac{1}{2} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \sum_{j=0}^{M} |w_j|^q$$
Ridge vs. Lasso

- Lasso tends to generate sparser solutions than a Ridge (quadratic) regularization
Ridge vs. Lasso

- Lasso tends to generate sparser solutions than a Ridge (quadratic) regularization (images taken from Derek Kane)

Ridge Regression  

Lasso Regression  

Least Absolute Shrinkage and Selection Operator
Figure 7.7  Degree 14 Polynomial fit to $N = 21$ data points with increasing amounts of $\ell_2$ regularization. Data was generated from noise with variance $\sigma^2 = 4$. The error bars, representing the noise variance $\sigma^2$, get wider as the fit gets smoother, since we are ascribing more of the data variation to the noise. Figure generated by linregPolyVsRegDemo.