Basic Regression Model and SVM Basis

1. Outline

- Linear Regression
    - Gradient Descent Methods
    - Underfitting vs. Overfitting: First Look
- Logistic Regression
- SVM
    - Optimal Margin Classifier
    - Generalized Lagrange Multiplier
    - Sequential Minimal Optimization for SVM

2 Linear Regression

2.1 Notations

- x： training data
- y： target variable or class label
- x^i：i-th data case
- x_j：j-yh feature of x
- m：total number of data case
- n：total number of feature

2.2 Model

Model Function： $h_\theta(x) = \theta_0 + \theta_1 x_1$

Let x_0 be a vector of 1s, $x_0 = [1, 1, 1, \ldots, 1]^T$: $h(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x$

The difference between target and predictions： $\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})$

Define the cost function： $J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (x^{(i)}) - y^{(i)})^2$

Our goal is to find the parameter that minimizing the cost function J(θ):

- $\hat{\theta} = \arg\min_\theta J(\theta)$
- One method: Gradient descent

2.3 Gradient Descent

- An iterative method to find the max/min values.
- Local max/min; sensitive to starting point
- Intuition: At each step, walk towards the steepest direction.
- Used widely in practice
- Steps:

1. Pick a starting point
2. Repeat
    a. Calculate the gradient of the function at current point
    b. Move a step towards the direction of the gradient to reduce J(θ)
3. Stop when J(θ) is small enough

- **Example of Gradient Descent in Linear Regression**

At each iteration, update the parameter $\theta_i \leftarrow \theta_i - \alpha\frac{\partial J(\theta)}{\partial \theta_i}$, where α is the step size.

$$
\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta_i} &= \frac{\partial}{\partial \theta_j}\frac{1}{2}\sum_{i=1}^{n}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 \\
&= \sum_{i=1}^{n}\left(h_\theta(x^{(i)}) - y^{(i)}\right)\frac{\partial}{\partial \theta_j}\left(h_\theta(x^{(i)}) - y^{(i)}\right) \\
&= \sum_{i=1}^{n}\left(h_\theta(x^{(i)}) - y^{(i)}\right)\frac{\partial}{\partial \theta_j}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right) \\
&Thus, \\
\frac{\partial J(\theta)}{\partial \theta_0} &= \sum_{i=1}^{n}\left(h_\theta(x^{(i)}) - y^{(i)}\right) \\
\frac{\partial J(\theta)}{\partial \theta_1} &= \sum_{i=1}^{n}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x^{(i)}
\end{aligned}
$$

2.4 Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to discriminative learning of linear classifiers under convex loss functions such as (linear) Support Vector Machines and Logistic Regression. Even though SGD has been around in the machine learning community for a long time, it has received a considerable amount of attention just recently in the context of large-scale learning. Stochastic Gradient Descent can be used:

Repeat:

- For i in 1 to m:

- Perform gradient calculation using only data point i

  $\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)}$

So far, we mentioned a iterative methods to calculate $\hat{\theta} = \arg\min_\theta J(\theta)$.

We can get more information from those topics and readings:

- Use matrix representation to calculate θ.
- The normal equations $\boldsymbol{\theta} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}} \underset{\boldsymbol{y}}{\rightarrow}$.
- http://cs229.stanford.edu/notes/cs229-notes1.pdf

  Part1, 2, Page 7-11.

2.5 Variation

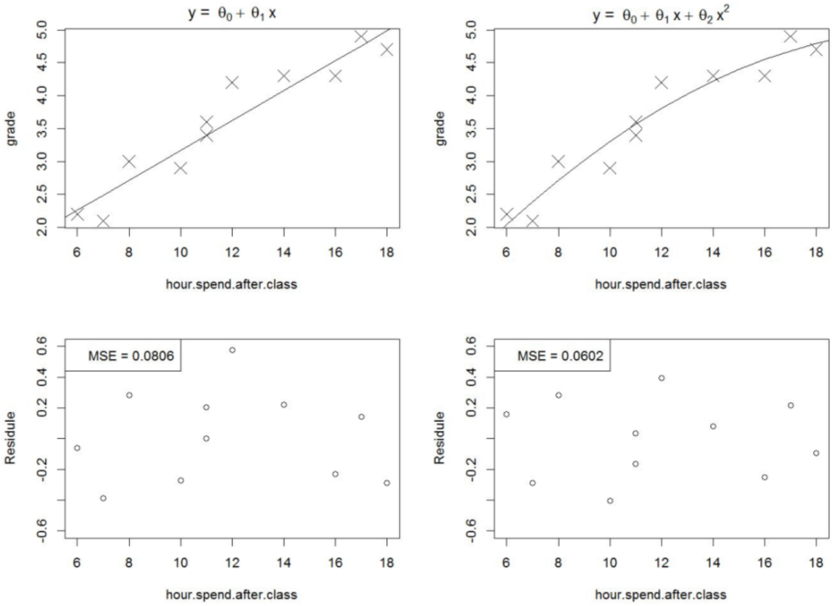
We can apply feature transformation to make the model appear “non-linear” .

EX: Polynomial Regression

- Useful when there is reason to believe the relationship between two variables is curvilinear.

- $y = \theta_0 + \theta_1 x + \theta_2 x^2$

- Might lead to overfitting.



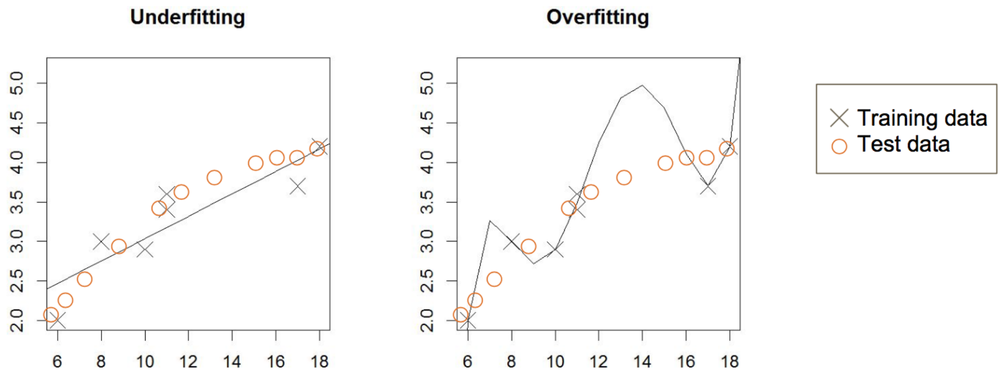
2.6 Bias vs. Variance (Underfitting vs. Overfitting)

Underfitting:

- Refers to a model that can neither model the training data nor generalize to new data.
- An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.

Overfitting:

- Overfitting refers to a model that models the training data too well.
- Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model.



3 Logistic Regression

- Logistic function: $\sigma(t) = \frac{e^t}{e^t+1} = \frac{1}{1+e^{-t}}$
- Logistic regression: $F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$, where F(x) is the probability of the target variable if X is "positive" : F(X) = P(Y = 1|X)
- It assumes the probability of the target variable being "positive" can be explained by a combination of the explanatory variable after logistic-transformation.
- Why is logistic regression related to linear regression?

  $$F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \Longrightarrow \frac{F(x)}{1-F(x)} = e^{\beta_0+\beta_1 x} \Longrightarrow \ln\left(\frac{F(x)}{1-F(x)}\right) = \beta_0 + \beta_1 x$$

  In $(\frac{F(x)}{1-F(x)})$ is called "odds". $(\frac{F(x)}{1-F(x)})$ is called "logit".
- In logistic regression, we assume a linear relationship between explanatory variable and the log odds of the target variable.

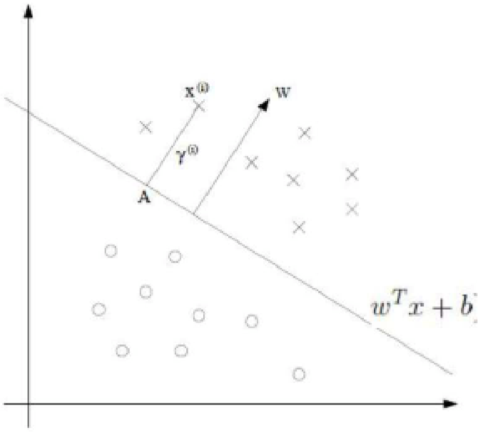
3.1 Other basic Models

- We not be talking about these models but they are worth looking at
  - Naive Bayes
  - K Nearest Neighbor
  - K-Means for clustering
  - Hierarchical clustering

4 Support Vector Machine Classifier (SVM)

Support vector machines (SVM) is a supervised learning model with associated learning algorithms that analyzes data used for both classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other categories, a SVM training algorithm can build a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

4.1 The Optimal Margin Classifier



As the graph shows, this is a linear classifier example. The distance from each point to classifier are:

$$\gamma^{(i)} = y^{(i)}\left(\left(\frac{w}{||w||}\right)^T x^{(i)} + \frac{b}{||w||}\right)$$

Note: The points in graph are $\gamma^{(\omega)}$, $x^{(\omega)}$, but the function use $\gamma^{(i)}$, $x^{(i)}$ to instead of.

If we want to gain the best hyper plane in classifier, we should find the maximize worst case distance:

$$\gamma = \min_{i=1,\dots,m} \gamma^{(i)}$$

Which after simplification, becomes:

$$\min_{\gamma,w,b} \frac{1}{2}||w||^2$$

$$s.t.\ y^{(i)}(w^T x^{(i)} + b) \geq 1,\ i = 1,\dots,m$$

formula deduction:

$$Decision\ Boundary(Hyper\ plane\ or\ classifier):\ w^T x + b = 0$$

$$A = x^{(i)} - \gamma^{(i)} \frac{w}{||w||}$$

$$\because A\ is\ on\ the\ boundary,$$

$$\therefore w^T(A) + b = 0$$

$$w^T\left(x^{(i)} - \gamma^{(i)} \frac{w}{||w||}\right) + b = 0$$

$$w^T x^{(i)} - \gamma^{(i)} \frac{w^T w}{||w||} + b = 0$$

$$\because w^T w = ||w||^2$$

$$\therefore \gamma^{(i)} ||w|| = w^T x^{(i)} + b$$

$$when\ x^{(i)} is\ positive, \gamma^{(i)} = \frac{w^T}{||w||} x^{(i)} + \frac{b}{||w||}$$

$$when\ x^{(i)} is\ negative, \gamma^{(i)} = -\frac{w^T}{||w||} x^{(i)} + \frac{b}{||w||}$$

$$Let\ y^{(i)} = \begin{cases} +1 & if\ x^{(i)} is\ positive \\ -1 & if\ x^{(i)} is\ negative \end{cases}\ Then: \gamma^{(i)} = y^{(i)}\left(\frac{w^T}{||w||} x^{(i)} + \frac{b}{||w||}\right)$$

Construct the optimization problem:

$$\gamma = \min \gamma^{(i)}$$

$$\max \gamma\ s.t.\ \ all\ points\ are\ outside\ of\ the\ margin\ \gamma$$

$$\max \frac{\gamma}{||w||}\ s.t.\ \ y^{(i)}\left(\frac{w^T}{||w||} x^{(i)} + \frac{b}{||w||}\right) \geq \frac{\gamma}{||w||}$$

$$Let\ \gamma = 1,$$

$$\max \frac{1}{||w||}\ s.t.\ \ y^{(i)}(w^T x^{(i)} + b) \geq 1$$

$$because\ ||w|| is\ a\ number,\ \ we\ have \min \frac{1}{2}||w||^2\ s.t.\ \ y^{(i)}(w^T x^{(i)} + b) \geq 1,\ for\ all\ i$$

4.2 Generalized Lagrange Multiplier – Primal Problem

For an optimization problem with inequality constraints:

$$\min_w f(w)$$
$$s.t.\ \ g_i(w) \leq 0, i = 1, \ldots, k$$
$$h_i(w) = 0, i = 1, \ldots, l$$

The Lagrangian (or generalized Lagrangian) is:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Define $\theta_{\mathcal{P}}(w) = \max_{\alpha,\beta:\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$

Then

$$\theta_{\mathcal{P}}(w) = \max_{\alpha,\beta:\alpha_i\geq 0} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\theta_{\mathcal{P}}(w) = \begin{cases} f(w) & if\ w\ satisfies\ primal\ constraints \\ \infty & otherwise \end{cases}$$

Thus $\min_w \theta_{\mathcal{P}}(w) = \min_w \max_{\alpha,\beta:\alpha_i\geq 0} \mathcal{L}(w,\alpha,\beta)$ has the same solution as the original problem.

4.3 Generalized Lagrange Multiplier – Dual Problem

Define$\theta_{\mathcal{D}}(\alpha,\beta) = \min_{\omega} \mathcal{L}(w,\alpha,\beta)$ and consider the dual problem

$$\max_{\alpha,\beta:\alpha_i\geq 0} \theta_{\mathcal{D}}(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i\geq 0} \min_{w} \mathcal{L}(w,\alpha,\beta)$$

Comparing the primal problem $\min_{w} \theta_{\mathcal{P}}(w) = \min_{w} \max_{\alpha,\beta:\alpha_i\geq 0} \mathcal{L}(w,\alpha,\beta)$

It can be shown that $d^* = \max_{\alpha,\beta:\alpha_i\geq 0} \min_{w} \mathcal{L}(w,\alpha,\beta) \leq \min_{w} \max_{\alpha,\beta:\alpha_i\geq 0} \mathcal{L}(w,\alpha,\beta) = p^*$

$d^* = p^*$ under certain conditions: 1. $g_i(w)$ are convex,

2. $h_i(w)$ can be written as $h_i(w) = a_i^T w + b_i$

Then there exist $w^*, \alpha^*, \beta^*\ s.t.\ p^* = d^* = \mathcal{L}(w^*, \alpha^*, \beta^*)$

4.4 Generalized Lagrange Multiplier – Primal/Dual Problem

If the above conditions are satisfied w^* ， α^* ， β^* also satisfy the KKT Dual complementarity conditions:

$$\frac{\partial}{\partial w_i}\mathcal{L}(w^*,\alpha^*,\beta^*) = 0, i = 1,\ldots,n$$

$$\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*,\alpha^*,\beta^*) = 0, i = 1,\ldots,l$$

$$\alpha_i^* g_i(w^*) = 0, i = 1,\ldots,k$$

$$g_i(w^*) leq 0, i = 1,\ldots,k$$

$$\alpha^* \geq 0, i = 1,\ldots,k$$

When $\alpha_i^* > 0$, (implies $g_i(w^*) = 0$), this constraint is said to be active.

Back to our optimization, $\min_{\gamma,w,b} \frac{1}{2}||w||^2\ s.t.\ y^{(i)}(w^T x^{(i)} + b) \geq 1, i = 1,\ldots,m$

After some derivation: (see next two slides for derivation)

$$\mathcal{L}(w,\alpha,b) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}$$

and the problem becomes:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)} x^{(j)} \rangle$$

$$s.t.\ \ \alpha_i \geq 0, i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

4.5 Sequential Minimal Optimization for SVM

The dual problem:

$$max_{\alpha}\, \mathrm{W}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} >$$

s.t $a_i \geq 0, \mathrm{i} = 1, \ldots, \mathrm{m}$

$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$

When updating the 1st variable, the constraints ensures that

$$\alpha_1 y^{(1)} = -\sum_{i=2}^{m} \alpha_i y^{(i)} \rightarrow \alpha_1 = -y^{(1)} \sum_{i=2}^{m} \alpha_i y^{(i)}$$

Update a pair of variable at one time.

```
Repeat till convergence{
    1. Select some pair $a_i$ and $a_j$ to update next (using a heuristic that tries to pick the two that will allow us to make the biggest progress towards the global maximum.
    2. Re-optimize W(α) with respect to $a_i$ and $a_j$, while holding all the other $a_k's$ $(k \neq i, j)$ fixed
}
```

4.6 Coordinate Ascent

(This method is used to solve the dual problem in the last slide)

Consider a maximization problem, $\max_{\alpha} W(\alpha_1, \alpha_2, \ldots, \alpha_m)$

The coordinate ascent algorithm:

```
Loop until convergence:{
    For i=1, ..., m,{
        $\alpha_i := \arg\max_{\hat{\alpha}_i} W(\alpha_1, \ldots, \alpha_{i-1}, \hat{\alpha}_i, \alpha_{i+1}, \ldots, \alpha_m)$
```

4.7 The Optimal Margin Classifier Review:

1. Problem formulation from geometry:

$$\min_{\gamma,w,b} \frac{1}{2} ||w||^2 \ s.t.\ \ y^{(i)}(w^T x^{(i)} + b) \geq 1, i = 1, \ldots, m$$

2. Construct Lagrangian:

$$\mathcal{L}(w,\alpha,b) = \frac{1}{2}||w||^2 - \sum_i \alpha_i \left(y^{(i)}(w^T x^{(i)} + b) - 1 \right)$$

3. Construct the dual problem:

$$d^* = \max_{\alpha} \min_{w,b} \frac{1}{2}||w||^2 - \sum_i \alpha_i \left(y^{(i)}(w^T x^{(i)} + b) - 1 \right)$$

4. Solver the inner optimization in the dual problem:

$$d^* = \max_{\alpha} \mathcal{L}^*(w,\alpha,b) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

5. Reformulate the problem:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$
$$s.t. \;\; \alpha_i \leq 0, i = 1, \ldots, m$$
$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

6. New problem to solve:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)} x^{(j)} \rangle$$
$$s.t. \;\; \alpha_i \geq 0, i = 1, \ldots, m$$
$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

7. After solving a_i by using Sequential Minimal Optimization:

$$w = \sum_i \alpha_i y^{(i)} x^{(i)}$$

$$b = \frac{1}{2}\left(\max_{i:y^{(i)}=-1} w^T x^{(i)} + \min_{i:y^{(i)}=1} w^T x^{(i)} \right)$$

8. To classify a new data case:

$$\begin{aligned} \hat{y}^{(i)} &= \phi(w^T x^{(i)} + b) \\ &= \phi\left(\sum_j \alpha_i y^{(i)} \langle x^{(i)}, x^{(j)} \rangle + b \right) \end{aligned}$$