CSE 417T: Introduction to Machine Learning

Lecture 8: Logistic Regression

Henry Chai
09/20/18
<table>
<thead>
<tr>
<th>Problem</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>$y = {-1, +1}$</td>
</tr>
<tr>
<td>Predicting Probabilities</td>
<td>$y = [0, 1]$</td>
</tr>
<tr>
<td>Regression</td>
<td>$y = \mathbb{R}$</td>
</tr>
<tr>
<td>Problem</td>
<td>Model</td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Linear Classification</td>
<td>$h(\hat{x}) = \text{sign}(\hat{w}^T \hat{x})$</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>$h(\hat{x}) = \theta(\hat{w}^T \hat{x})$</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>$h(\hat{x}) = \hat{w}^T \hat{x}$</td>
</tr>
</tbody>
</table>
Predicting Probabilities

- Training data **does not** consist of probabilities
- Observations are still **binary**: $y_i = \pm 1$
- Goal is to learn $f(\vec{x}) = P\{y = +1|\vec{x}\}$
- Observations are inherently noisy
\[ h(\hat{x}) = P\{y = +1|\hat{x}\} \]

\[ h(\hat{x}) = \theta(\overline{w}^T\hat{x}) = \frac{1}{1+e^{-\overline{w}^T\hat{x}}} = \left(1 + e^{-\overline{w}^T\hat{x}}\right)^{-1} \in [0,1] \]

Note that \( 1 - \theta(\overline{w}^T\hat{x}) = \theta(-\overline{w}^T\hat{x}) \)

\[ 1 - \theta(\overline{w}^T\hat{x}) = 1 - \frac{1}{1+e^{-\overline{w}^T\hat{x}}} = \frac{e^{-\overline{w}^T\hat{x}}}{1+e^{-\overline{w}^T\hat{x}}} = \frac{1}{1+e^{\overline{w}^T\hat{x}}} = \theta(-\overline{w}^T\hat{x}) \]
This error metric is a viable choice that behaves how we want but is difficult to optimize.
The cross-entropy error looks weird but actually arises naturally from our choice of hypothesis set and is also easier to minimize.
The key is to assume conditional independence

\[ P(D|h) = P\{(\mathbf{x}_1, y_1) \cap \cdots \cap (\mathbf{x}_n, y_n) | h\} \]
\[ = \prod_{i=1}^{n} P\{(\mathbf{x}_i, y_i) | h\} \]
\[ = \left( \prod_{i: y_i = +1} \theta(\mathbf{w}^T \mathbf{x}_i) \right) \left( \prod_{i: y_i = -1} 1 - \theta(\mathbf{w}^T \mathbf{x}_i) \right) \]
\[ = \left( \prod_{i: y_i = +1} \theta(\mathbf{w}^T \mathbf{x}_i) \right) \left( \prod_{i: y_i = -1} \theta(-\mathbf{w}^T \mathbf{x}_i) \right) \]
\[ = \prod_{i=1}^{n} \theta(y_i \mathbf{w}^T \mathbf{x}_i) = \prod_{i=1}^{n} \left( 1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i} \right)^{-1} \]
Find $\hat{w}^*$ that maximizes $P(D|\hat{w}) = \prod_{i=1}^{n} \left( 1 + e^{-y_i \hat{w}^T x_i} \right)^{-1}$

Find $\hat{w}^*$ that maximizes $P(D|\hat{w}) = \ln \left( \prod_{i=1}^{n} \left( 1 + e^{-y_i \hat{w}^T x_i} \right)^{-1} \right)$

Find $\hat{w}^*$ that maximizes $P(D|\hat{w}) = \sum_{i=1}^{n} \ln \left( 1 + e^{-y_i \hat{w}^T x_i} \right)$

Find $\hat{w}^*$ that maximizes $P(D|\hat{w}) = -\sum_{i=1}^{n} \ln \left( 1 + e^{-y_i \hat{w}^T x_i} \right)$

Find $\hat{w}^*$ that **minimizes** $P(D|\hat{w}) = \frac{1}{n} \sum_{i=1}^{n} \ln \left( 1 + e^{-y_i \hat{w}^T x_i} \right)$
So unfortunately, even though this is the “right” error metric to use, setting the gradient w.r.t. $\bar{w}$ equal to 0 and solving for the optimal $\bar{w}$ is impossible.
So unfortunately, even though this is the “right” error metric to use; setting the gradient w.r.t. $\vec{w}$ equal to 0 and solving for the optimal $\vec{w}$ is impossible.
So unfortunately, even though this is the “right” error metric to use; setting the gradient w.r.t. $\vec{w}$ equal to 0 and solving for the optimal $\vec{w}$ is impossible.
Gradient Descent: Intuition

- Iterative method for minimizing functions
- Requires the gradient to exist everywhere
Functions where there is only 1 global minimum are called convex. In 1d, such functions are characterized by having non-negative second-order derivatives (think $x^2$). The corollary in higher dimensions is the Hessian matrix being positive semidefinite.
Cross-entropy Error

\[ E_{\text{in}}(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} \ln \left( 1 + e^{-y_i \sum_{j=0}^{d} w_j x_{ij}} \right) \]

\[ \frac{\partial E_{\text{in}}(\vec{w})}{\partial w_j} = \frac{1}{n} \sum_{i=1}^{n} \frac{-y_i x_{ij} e^{-y_i \sum_{j=0}^{d} w_j x_{ij}}}{1 + e^{-y_i \sum_{j=0}^{d} w_j x_{ij}}} \]

\[ \nabla_{\vec{w}} E_{\text{in}}(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} \frac{-y_i x_{i}^T}{e^{y_i \vec{w}^T \vec{x}_i} + 1} \]

\[ H_{\vec{w}} E_{\text{in}}(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i^2 e^{y_i \vec{w}^T \vec{x}_i}) (\vec{x}_i \vec{x}_i^T)}{(e^{y_i \vec{w}^T \vec{x}_i} + 1)^2} \]

\( H_{\vec{w}} E_{\text{in}}(\vec{w}) \) is positive semidefinite \( \forall \vec{w} \rightarrow E_{\text{in}}(\vec{w}) \) is convex
Gradient Descent: Intuition

- $w^{(t+1)} = w^{(t)} + \eta \nabla f(w^{(t)})$
- Suppose the current location is $w^{(t)}$
  - Move some distance $\nabla f(w^{(t)})$ in the "most downhill" direction possible.
\( \hat{v} \)

- Fix \( \eta \) and choose \( \hat{v} \) to maximize the decrease in \( E_{in} \) after making the update \( \tilde{w}_{(t+1)} = \tilde{w}_{(t)} + \eta \hat{v} \)

- \( \Delta E_{in} = E_{in}(\tilde{w}_{(t)} + \eta \hat{v}) - E_{in}(\tilde{w}_{(t)}) \)
\[ \hat{\nu} \]

- Fix \( \eta \) and choose \( \hat{\nu} \) to minimize \( \Delta E_{in} \) after making the update \( \overline{w}_{(t+1)} = \overline{w}_{(t)} + \eta \hat{\nu} \)

\[ \Delta E_{in}(\hat{\nu}) = E_{in}(\overline{w}_{(t)} + \eta \hat{\nu}) - E_{in}(\overline{w}_{(t)}) \]
Multivariate first-order Taylor series expansion:

\[ f(\tilde{x}) \approx f(\tilde{a}) + (x - a)^T \nabla f(\tilde{a}) \]

The fourth line follows because \( \tilde{v} \) is a unit vector and the inner product of a unit vector and any arbitrary vector is at most the L2 norm of the non-unit vector and is at least \(-1 \cdot \text{L2 norm of the non-unit vector}\) (depending on the direction the unit vector points in).
\[ \eta \]

Small \( \eta \)

Large \( \eta \)
- Use a variable $\eta_t$ instead of a fixed $\eta$
• Set $\eta_t = \eta_0 \| \nabla_{\bar{w}} E_{in}(\bar{w}_t) \|$ 

• $\eta_t$ decreases as $t$ increases, because $\| \nabla_{\bar{w}} E_{in}(\bar{w}_t) \|$ decreases as $E_{in}(\bar{w}_t)$ approaches its minimum 

• $\bar{w}_{(t+1)} = \bar{w}_{(t)} + \eta_t \bar{v}_t^*$ 
  $\quad = \bar{w}_{(t)} + (\eta_0 \| \nabla_{\bar{w}} E_{in}(\bar{w}_t) \|) \left( - \frac{\nabla_{\bar{w}} E_{in}(\bar{w}_t)}{\| \nabla_{\bar{w}} E_{in}(\bar{w}_t) \|} \right)$ 
  $\quad = \bar{w}_{(t)} - \eta_0 \nabla_{\bar{w}} E_{in}(\bar{w}_t)$