Large Scale Optimization for Machine Learning

Meisam Razaviyayn
Lecture 11
razaviya@usc.edu
Agenda

• Some optimization problems in ML:
  • SVM, regression, logistic regression, deep learning

• Empirical risk minimization framework
  • Generalization error, tradeoffs, cross validation

• Exploiting structure
### Linear Regression

<table>
<thead>
<tr>
<th>Area</th>
<th>Crime Rate</th>
<th>Age</th>
<th>RAD</th>
<th>PTRATIO</th>
<th>Bedrooms</th>
<th>Price (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>600</td>
<td>1.05</td>
<td>12</td>
<td>2.4</td>
<td>10.1</td>
<td>1</td>
<td>500</td>
</tr>
<tr>
<td>1000</td>
<td>2.34</td>
<td>10</td>
<td>2.5</td>
<td>20.1</td>
<td>1</td>
<td>800</td>
</tr>
<tr>
<td>1200</td>
<td>1.45</td>
<td>3</td>
<td>3.1</td>
<td>9.7</td>
<td>3</td>
<td>1500</td>
</tr>
<tr>
<td>1500</td>
<td>1.56</td>
<td>30</td>
<td>1.7</td>
<td>7.2</td>
<td>2</td>
<td>1200</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>2700</td>
<td>1.01</td>
<td>20</td>
<td>0.9</td>
<td>4.3</td>
<td>4</td>
<td>5000</td>
</tr>
</tbody>
</table>

\[ y = w^T x + z \text{ with } z \sim \mathcal{N}(0, \sigma^2) \]

Maximum Likelihood Estimation

\[ \min_{w} \frac{1}{2} \sum_{i=1}^{n} (w^T x_i - y_i)^2 \]

s.t. \( w \in \mathbb{R}^d \)

What is the loss function here?
Logistic Regression

<table>
<thead>
<tr>
<th>Radius</th>
<th>Texture</th>
<th>Area</th>
<th>Compactness</th>
<th>Symmetry</th>
<th>...</th>
<th>Rec/non-Rec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>2.3</td>
<td>3.5</td>
<td>2.4</td>
<td>1.4</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>0.7</td>
<td>1.2</td>
<td>2.5</td>
<td>1.4</td>
<td>3.2</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>1.7</td>
<td>2.4</td>
<td>1.5</td>
<td>3.3</td>
<td>1.3</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>0.2</td>
<td>3.4</td>
<td>0.7</td>
<td>4.3</td>
<td>2.0</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>

Model: \( \log \left( \frac{P(y = 1 \mid w, x)}{P(y = 0 \mid w, x)} \right) = w^T x \)

Maximum likelihood estimator

\[
\min_w \sum_{i=1}^{n} \log (1 + \exp(w^T x_i)) - \sum_{\{i : y_i = 1\}} w^T x_i
\]

s.t. \( w \in \mathbb{R}^d \)
Logistic Regression

Model: \[ \log \left( \frac{\mathbb{P}(y = 1 \mid \mathbf{w}, \mathbf{x})}{\mathbb{P}(y = 0 \mid \mathbf{w}, \mathbf{x})} \right) = \mathbf{w}^T \mathbf{x} \]

Maximum likelihood estimator

What is the loss function here?

\[ \min_{\mathbf{w}} \sum_{i=1}^{n} \log (1 + \exp(\mathbf{w}^T \mathbf{x}_i)) - \sum_{\{i: y_i = 1\}} \mathbf{w}^T \mathbf{x}_i \]

s.t. \( \mathbf{w} \in \mathbb{R}^d \)

\[ \min_{\mathbf{w}} \sum_{i=1}^{n} \left( \log (1 + \exp(\mathbf{w}^T \mathbf{x}_i)) - y_i \mathbf{w}^T \mathbf{x}_i \right) \]
Support Vector Machines

• Binary classification task

<table>
<thead>
<tr>
<th>Blood Pressure</th>
<th>Age</th>
<th>Sex</th>
<th>LDL</th>
<th>Glucose</th>
<th>BMI</th>
<th>...</th>
<th>Diabetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>85</td>
<td>29</td>
<td>0</td>
<td>100</td>
<td>75.1</td>
<td>24.2</td>
<td>...</td>
<td>-1</td>
</tr>
<tr>
<td>95</td>
<td>50</td>
<td>1</td>
<td>115</td>
<td>90.2</td>
<td>19.2</td>
<td>...</td>
<td>-1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>123</td>
<td>42</td>
<td>1</td>
<td>150</td>
<td>110</td>
<td>25.2</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>

New patient:

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>119</td>
<td>37</td>
<td>0</td>
<td>120</td>
<td>100</td>
<td>19.3</td>
<td>...</td>
<td>???</td>
</tr>
</tbody>
</table>

**Task:** find a mapping $h : \text{Features} \mapsto \text{Labels}$
Support Vector Machines

Which one is better? → Maximum margin classifier
Support Vector Machines

\[ \max \quad \frac{2}{\|w\|} \]
\[ \text{s.t.} \quad y_i(w^T x_i + v) \geq 1 \]

Convex optimization

\[ \min_{w,v} \quad \|w\|_2^2 \]
\[ \text{s.t.} \quad y_i(w^T x_i + v) \geq 1 \]
Support Vector Machines: Soft-Margin

Might be infeasible

\[
\min_{w,v} \quad \|w\|_2^2 \\
\text{s.t.} \quad y_i(w^T x_i + v) \geq 1
\]

Soft-Margin SVM:

\[
\min_{w,v} \quad \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i(w^T x_i + v)\} + \lambda \|w\|_2^2
\]

What is the loss function here?
Neural Networks: Digit classification

Should be able to construct complex predictors

[Deep learning, Ian Goodfellow et al. 2016]
SVM, regression, logistic regression can be viewed as neural networks with one layer.
Empirical Risk Minimization Framework

Predicting an output $y \in \mathcal{Y}$ given an input $x \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} \in \{0, 1\}$

Set of hypotheses: $\mathcal{H}$ with $h \in \mathcal{H}$ maps $\mathcal{X}$ to $\mathcal{Y}$

Loss function: $\ell : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{H} \mapsto \mathbb{R}$

Data generating distribution $\mathbb{P}^*$ with $(x, y) \sim \mathbb{P}^*$

**Expected risk/Test error:** $L(h) \triangleq \mathbb{E}_{\mathbb{P}^*}[\ell((x, y), h)]$

Training samples: $(x_1, y_1), \ldots, (x_n, y_n)$

**Empirical risk/Training error:** $\hat{L}(h) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell((x_i, y_i), h)$

Expected risk of ERM: $L(\hat{h})$

Best we can hope for

Empirical Risk Minimizer
What Set of Hypotheses to Consider?

Simple $\mathcal{H}$

Underfitting

Complex $\mathcal{H}$

Overfitting

Trade-offs between “number of samples”, “Expected risk or ERM”, “Complexity of hypothesis class”
There are different ways of measuring complexity of a hypothesis class, but in general this trade-off exists...
Simple Case

Assume:

\[ L(h^*) = \mathbb{E}_{P^*} [\ell((x, y), h^*)] = 0 \]

|\mathcal{H}| < \infty

zero-one loss: \( \ell((x, y), h) = \mathbb{I}[y \neq h(x)] \)

Then, with probability at least 1 − \( \delta \)

\[ L(\hat{h}) - L(h^*) \leq \frac{\log |\mathcal{H}| + \log(1/\delta)}{n} \]

Proof?

Equivalent statement provides sample complexity
Remarks

Regularization typically reduces model complexity

How to estimate expected risk? Cross Validation!

Structure of the problems

- Summation in the objective
- Stochastic optimization
- Online optimization
- Incremental methods

- Large number of blocks
- Block methods

\[
\min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2
\]

\[
\min_w \sum_{i=1}^{n} (\log (1 + \exp(w^T x_i)) - y_i w^T x_i)
\]

\[
\min_{w,v} \frac{1}{n} \sum_{i=1}^{n} \max \{ 0, 1 - y_i (w^T x_i + v) \} + \lambda ||w||^2
\]