10: Empirical Risk Minimization

Remember the unconstrained SVM Formulation

\[
\min_w C \sum_{i=1}^n \max \left[ 1 - y_i \left( w^\top x_i + b \right), 0 \right] + \frac{\|w\|^2}{2} \tag{Hinge-Loss}
\]

The hinge loss is the SVM's error function of choice, whereas the \(l_2\)-regularizer penalizes (overly) complex solutions. This is an example of empirical risk minimization with a loss function \(\ell\) and a regularizer \(r\),

\[
\min_w \frac{1}{n} \sum_{i=1}^n \ell(h_w(x_i), y_i) + \lambda r(w)
\]

where the loss function is a continuous function which penalizes training error, and the regularizer is a continuous function which penalizes classifier complexity.

Here, we define \(\lambda\) as \(\frac{1}{C}\) from the previous lecture.\(^{[1]}\)

Commonly Used Binary Classification Loss Functions

Different Machine Learning algorithms use different loss functions; Table 4.1 shows just a few (here we assume \(y_i \in \{+1, -1\}\)):

<table>
<thead>
<tr>
<th>Loss (\ell(h_w(x_i), y_i))</th>
<th>Usage</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hinge-Loss</strong> (\max \left[ 1 - h_w(x_i) y_i, 0 \right]^p)</td>
<td>• Standard SVM ((p = 1)) • (Differentiable) Squared</td>
<td>When used for Standard SVM, the loss function denotes the size of the margin between linear separator and its closest points in either class. Only</td>
</tr>
</tbody>
</table>
### Table 4.1: Loss Functions With Classification $y \in \{-1, +1\}$

<table>
<thead>
<tr>
<th>Loss Function</th>
<th>Formula</th>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Log-Loss</strong></td>
<td>$\log(1 + e^{-h_w(x_i)y_i})$</td>
<td>Logistic Regression</td>
<td>One of the most popular loss functions in Machine Learning, since its outputs are well-calibrated probabilities.</td>
</tr>
<tr>
<td><strong>Exponential Loss</strong></td>
<td>$e^{-h_w(x_i)y_i}$</td>
<td>AdaBoost</td>
<td>This function is very aggressive. The loss of a mis-prediction increases <em>exponentially</em> with the value of $-h_w(x_i)y_i$. This can lead to nice convergence results, for example in the case of AdaBoost, but it can also cause problems with noisy data.</td>
</tr>
<tr>
<td><strong>Zero-One Loss</strong></td>
<td>$\delta(\text{sign}(h_w(x_i)) \neq y_i)$</td>
<td>Actual Classification Loss</td>
<td>Non-continuous and thus impractical to optimize.</td>
</tr>
</tbody>
</table>

**Quiz:** What do all these loss functions look like with respect to $z = yh(x)$? Some questions about the loss functions:

1. Which functions are strict upper bounds on the $0/1$-loss?
2. What can you say about the hinge-loss and the log-loss as $z \to -\infty$?

Some additional notes on loss functions:

3. As $z \to -\infty$, the log-loss and the hinge loss become increasingly parallel.
4. The exponential loss and the hinge loss are both upper bounds of the zero-one loss. (For the exponential loss, this is an important aspect in Adaboost, which we will cover later.)
5. Zero-one loss is zero when the prediction is correct, and one when incorrect.
Commonly Used Regression Loss Functions

Regression algorithms (where a prediction can lie anywhere on the real-number line) also have their own host of loss functions:

<table>
<thead>
<tr>
<th>Loss $\ell(h_w(x_i, y_i))$</th>
<th>Comments</th>
</tr>
</thead>
</table>
| **Squared Loss** $(h(x_i) - y_i)^2$ | • Most popular regression loss function  
• Estimates [Mean](#) Label  
• Also known as [Ordinary Least Squares](#) (OLS)  
• 😊 Differentiable everywhere  
• 😞 Somewhat sensitive to outliers/noise |
| **Absolute Loss** $|h(x_i) - y_i|$ | • Also a very popular loss function  
• Estimates [Median](#) Label  
• 😊 Less sensitive to noise  
• 😞 Not differentiable at 0 |
| **Huber Loss**  
• $\frac{1}{2}(h(x_i) - y_i)^2$ if $|h(x_i) - y_i| < \delta$,  
• otherwise  
$\delta(|h(x_i) - y_i| - \frac{\delta}{2})$ | • Also known as [Smooth Absolute Loss](#)  
• "Best of Both Worlds" of [Squared](#) and [Absolute](#) Loss  
• Once-differentiable  
• Takes on behavior of Squared-Loss when loss is small, and Absolute Loss when loss is large. |
| **Log-Cosh Loss**  
$log(cosh(h(x_i) - y_i))$,  
$cosh(x) = \frac{e^x + e^{-x}}{2}$ | 😊 Similar to Huber Loss, but twice differentiable everywhere  
• 😞 More expensive to compute |

Table 4.2: Loss Functions With Regression, i.e. $y \in \mathbb{R}$

**Quiz:** What do the loss functions in Table 4.2 look like with respect to $z = h(x_i) - y_i$?
Regularizers

When we investigate regularizers it helps to change the formulation of the optimization problem from an unconstrained to a constraint formulation, to obtain a better geometric intuition:

\[
\min_{w,b} \sum_{i=1}^{n} \ell(h_w(x), y_i) + \lambda r(w) \Leftrightarrow \min_{w,b} \sum_{i=1}^{n} \ell(h_w(x), y_i) \text{ subject to: } r(w) \leq B
\]

For each \( \lambda \geq 0 \), there exists \( B \geq 0 \) such that the two formulations above are equivalent, and vice versa. In previous sections, we have already seen the \( l_2 \)-regularizer in the context of SVMs, Ridge Regression, or Logistic Regression. Besides the \( l_2 \)-regularizer, other types of useful regularizers and their properties are listed in Table 4.3.

<table>
<thead>
<tr>
<th>Regularizer ( r(w) )</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>( l_2 )-Regularization ( r(w) = w^T w = |w|_2^2 )</td>
<td>☺ Strictly Convex&lt;br&gt;☺ Differentiable&lt;br&gt;=./ Uses weights on all features, i.e. relies on all features to some degree (ideally we would like to avoid this) - these are known as Dense Solutions.</td>
</tr>
<tr>
<td>( l_1 )-Regularization ( r(w) =</td>
<td>w</td>
</tr>
<tr>
<td>( l_p )-Norm ( |w|<em>p = (\sum</em>{i=1}^{d}</td>
<td>v_i</td>
</tr>
</tbody>
</table>

Table 4.3: Most popular Regularizers

Famous Special Cases

This section includes several special cases that deal with risk minimization, such as Ordinary Least Squares, Ridge Regression, Lasso, and Logistic
Regression. Table 4.4 provides information on their loss functions, regularizers, as well as solutions.

<table>
<thead>
<tr>
<th>Loss and Regularizer</th>
<th>Comments</th>
</tr>
</thead>
</table>
| **Ordinary Least Squares** | - Squared Loss  
- No Regularization  
- Closed form solution:  
  \[ w = (XX^\top)^{-1}Xy^\top \]  
  \[ X = [x_1, \ldots, x_n] \]  
  \[ y = [y_1, \ldots, y_n] \] |
| **Ridge Regression** | - Squared Loss  
- \( l_2 \)-Regularization  
- \( w = (XX^\top + \lambda I)^{-1}Xy^\top \) |
| **Lasso** | - \( \leq \) sparsity inducing (good for feature selection)  
- \( \leq \) Convex  
- \( \leq \) Not strictly convex (no unique solution)  
- \( \leq \) Not differentiable (at 0)  
- Solve with (sub)-gradient descent or SVEN |
| **Elastic Net** | - \( \leq \) Strictly convex (i.e. unique solution)  
- \( \leq \) sparsity inducing (good for feature selection)  
- \( \leq \) Dual of squared-loss SVM, see SVEN  
- \( \leq \) Non-differentiable |
| **Logistic Regression** | - Often \( l_1 \) or \( l_2 \) Regularized  
- Solve with gradient descent.  
- \[ P_r(y|x) = \frac{1}{1+e^{-y(w^\top x+b)}} \] |
Linear Support Vector Machine

\[ \min_{\mathbf{w}, b} C \sum_{i=1}^{n} \max[1 - y_i (\mathbf{w}^\top \mathbf{x}_i + b), 0] + \|\mathbf{w}\|_2^2 \]

- Typically \( l_2 \) regularized (sometimes \( l_1 \)).
- Quadratic program.
- When kernelized leads to sparse solutions.
- Kernelized version can be solved very efficiently with specialized algorithms (e.g. SMO)

Table 4.4: Special Cases

Some additional notes on the Special Cases:

1. Ridge Regression is very fast and can be solved in closed form if the data isn't too high dimensional (in just 1 line of code.)
2. There is an interesting connection between Ordinary Least Squares and the first principal component of PCA (Principal Component Analysis). PCA also minimizes square loss, but looks at perpendicular loss (the horizontal distance between each point and the regression line) instead.

[1] In Bayesian Machine Learning, it is common to optimize \( \lambda \), but for the purposes of this class, it is assumed to be fixed.