CHEAT SHEET

Regularizers

When you look at regularizers, it helps to change the formulation of the optimization problem to obtain a better geometric intuition:

$$\min_{\mathbf{w},b} \sum_{i=1}^{n} \ell(h_{\mathbf{w}}(\mathbf{x}_i)) + \lambda r(\mathbf{w}) \Longleftrightarrow \min_{\mathbf{w},b} \sum_{i=1}^{n} \ell(h_{\mathbf{w}}(\mathbf{x}_i))$$

subject to: $r(\mathbf{w}) \leq B$

Regularizers		Details
l_2 -Regularization	$r(\mathbf{w}) = \mathbf{w}^\top \mathbf{w} = \|\|\mathbf{w}\|\|_2^2 = \sum_{i=1}^{d} [\mathbf{w}]_i^2$	• ADVANTAGE: Strictly Convex • ADVANTAGE: Differentiable • DISADVANTAGE: 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.
l_1 -Regularization	$r(\mathbf{w}) = \|\|\mathbf{w}\|\|_1 = \sum_{i=1}^{d} \|[\mathbf{w}]_i\|$	• Convex (but not strictly) • DISADVANTAGE: Not differentiable at 0 • Effect: Sparse
l_p -Norm	$r(\mathbf{w}) = \|\|\mathbf{w}\|\|_p^p = d\sum_{i=1}^{d} [\mathbf{w}]_i^p$	• Often $0 < p \leq 1$ • DISADVANTAGE: Non-convex • ADVANTAGE: Very sparse solutions • Initialization dependent • VERY sparse solutions (compared to l1 norm) if $0 < p \leq 1$.