Regularization in Neural Networks

Justin Talbot, James Loomis, Wenbo Wang
Regularization

• Bias / Variance Trade off [1]

\[ B(D) = E[g(z; D) - I(z)] \]

\[ V(D) = E[(g(z; D) - E[g(z; D)])^2] \]

• Increasing bias will result in decreasing variance, and vice versa

[1] Justin, Domke, Lecture note of Statistical Machine Learning, 2009 Spring Quarter, RIT
Bias / Variance Trade off

[1] Justin, Domke, Lecture note of Statistical Machine Learning, 2009 Spring Quarter, RIT
Bayesian Interpretation

• Given the prior density function of the weights $p(w)$

• The error function

$$E = E(D) + \lambda E_w(w)$$

could be interpreted with conditional probability

$$P(w | D) \propto P(D | w) P(w)$$
Gaussian Weight Prior

\[ \tilde{E}(w) = E(w) + \lambda \| w \|^2 \]
Laplacian Weight Prior

\[ \tilde{E}(w) = E(w) + \lambda |w| \]

- Induces parameter sparsity by shrinking variables to exactly 0
- Non-differentiable

\[ \text{ } \]
Elastic Net Prior

\[ \tilde{E}(w) = E(w) + \alpha (w^T w) + (1 - \alpha) (|w|) \]

- Parameters go exactly to zero like in Laplacian priors but highly correlated variables are shrunk together like in Gaussian priors.

Early Stopping

• To end the training early before it converges
• Learning in NN is a highly non-convex optimization
• It is unclear exactly how early stopping will affect the objective function
• Under a quadratic error function early stopping is approximately a Gaussian prior
Invariances

• Regularization by making the classifier invariant to transformations of the data
• Can be done by adjusting the training data directly or the classifier directly
Convolutional Networks
Soft Weight Sharing

- The prior probability is a mixture of gaussians
- Pushes the weight values to form several groups
- The weights in the group tend towards the same value