Paper: Robert Tibshirani, "Regression Shrinkage and Selection Via Lasso," Journal of the Royal Statistical Society, 1996.

Summary: The author introduces the lasso by comparing it with other learning models on qualities of good models. The set-up, motivation and its geometric interpretation of the lasso with its coefficient shrinkage compared with other models are given. of the lasso are given. The author further elaborates on two special cases of the lasso and discusses the difficulties and approaches to estimate its standard error. Example fits on prostate cancer data using the lasso and other techniques are compared in their model coefficient estimates results. To tune the regularization parameter to reduce the model error, three methods are derived and their restrictions and processing times are compared and this tuning problem is reframed in a Bayesian likelihood setting. The author moves on to explain an algorithm solving the lasso optimization problem and provides motivation, testing, and extensions of it. The strengths and weaknesses of the lasso and other models are examined by using four simulations with differing observations and parameter count and true parameter values. Application of the lasso are shown in regression models and model pruning. The author compares soft and hard thresholding of the coefficients and summarizes the performance and properties of the lasso and alternative models.

Two Uses or Extensions

- The lasso method allows some coefficients so be 0 so the equation is easier to interpret because only the important features are highlighted. However, the method does not give an indication on the correlation between pairs of the features with non-zero coefficients so a pair of features with high correlation can be simplified into a single feature with a weight approximately the sum of the two smaller weights, which would make the model easier to interpret. To integrate this change, the covariance matrix for the features can be computed and if a feature has high covariance with many other features, then it will have a less likely chance of having a non-zero coefficient.

- The lasso method can be used in computer vision to associate a person's face with a property of that person such as income. There are a moderately large number of independent features on a person's face (e.g. measurements of eyes, mouth, ears, and nose) and the independence of these features implies their effects are unlikely to be concentrated to either all be high or low due to similar measurement magnitudes. The moderate nature of both the quantity and effects of features makes the lasso method well-suited to provide an equation for the property based on these features.

Paper: Leon Bottou and Olivier Bousquet, "The Tradeoffs of Large Scale Learning," Advances in Neural Information Processing Systems (NeurIPS), 2007.

Summary: The authors begin by reinforcing the premise that time complexity is an important but overlooked aspect of learning models and argue for the necessity and feasibility of such models whose size grows at equal rates with the total data inputted. The authors introduces optimization error and explains its importance in large models and implications from using it. The authors first define the approximation and estimation error associated with a sample probability distribution, loss function, and estimator universe and give the interpretations and properties of these errors. The authors then extend the equations with the optimization error and interpreted it as an estimator function's generalizability. The authors relate the three errors and time complexity with model parameters and divide small and large models using time and input size. The authors restrict large models with justification and bound the errors and comment on their convergence rates. The authors define calculus concepts to analyze the recursions and time complexity of first and second order stochastic and standard gradient descent and compare their accuracy and time complexity to justify the importance of the optimization error. The authors end by highlighting analysis parameters that can be changed for further work.

Two Uses or Extensions

- The authors make a binary distinction between small and large scale problems by defining small scale problems to be those where computing time is not a constraint and large scale as all the other problems. In practice, less computing time is always desirable so one way to capture this is to introduce a continuous variable α indicating the importance of computing time where a problem with short computing times have smaller α. This removes the two cases of small and large scale problems and the analysis in table 2 can be performed on all problems. There can be an additional "practical score" for each algorithm involving α indicating how practical the algorithm is given its generalization ability, accuracy, and importance of computing time α.
- The assumption $E_n(\tilde{f}_n) < E_n(f_n) + \rho$ does not apply to cases where the approximate solution $\tilde{f}_n$ can have smaller error when the magnitude of the $E_n(f_n)$ is smaller. In this case, a value of ρ would not exist because the error can grow arbitrarily large as $E_n(f_n)$ increases. The assumption can be generalized to $E_n(\tilde{f}_n) < g(E_n(f_n))$ for some function g and with the special linear case $E_n(\tilde{f}_n) < \beta E_n(f_n) + \rho$ for some approximation factor β can have its convergence analyzed using extension of the bounds used in the paper.

Paper: Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems (NeurIPS), 2012.

Summary: The authors state the need of large datasets for image classification and examine the strengths and weaknesses involving efficieny of convolutional neural networks. The authors inform the reader of the results, features, and next steps of their model before explaining the ImageNet dataset, the performance metrics, and transformations of the dataset before using it as input. The authors then have a discussion of specific methods used in the network, including a argument of using of ReLU layer over other activation functions, the organization and independence of the two GPUs used, the rationale and background of a normalization technique, and the benefits of sampling overlapping square sub-grids. The authors provide an overview interaction and activations of the layers used, including convolutional, fully-connected, normalization, and sampling layers. The authors discuss the need to reduce overfitting and techniques used such as generating sub-images, colour vector manipulation, and randomly eliminating neuron outputs. The authors then discuss the specifics of the stochastic gradient descent used such as its recursion, initialization, and learning trajectory. The authors discuss the competition results of their model against other models, their model's thought process and properties, and importance and extensions through scaling of their model's depth and data volume.

Two Uses or Extensions

- In addition to the five images generated from the first data augmenting technique, an image can be generated where the RGB values of a pixel is made more intense depending on values of the up to eight neighbouring pixels. If many of these eight pixels have similar RGB value of the center pixel, then the center pixel will have its colour highlighted since it is likely to be part of the foreground or the background. This will create a new image from the original where the foreground and background have greater contrast so the network can learn the features of the foreground more easily and associate it with the label.

- The dropout technique can be used in k-nearest neighbours using m features where when predicting a point A, a subset of size $n < m$ of features and the k closest neighbours of A will be computed using the size n subset. This is helpful because the Euclidean distance metric and many other distance metrics magnify the importance of features with high variance so they tend to be the most important in selecting the k nearest neighbours. By randomizing the selection, such features will have a probability less than 1 of being used in the computation, which makes the used features have more equal effects. Similar to the dropout technique, this will reduce overfitting because the random nature of the feature subset selected causes the model to not learn any specific subset of features that previously would have the greatest predictive power in the training set.

Paper: Moritz Hardt, Eric Price, and Nathan Srebro, "Equality of opportunity in supervised learning," Advances in Neural Information Processing Systems (NeurIPS), 2016.

Summary: The authors emphasize the need for discrimination-free machine learning models and the scarcity of such models. The authors argue why ignoring protected attributes is not ideal in fairness and utility and create a supervised learning setting while noting the non-necessity of predictor derivation. The authors overview their fairness concept using comparisons with other such concepts and give viability reasoning. The authors define equalized odds, give an intepretation in terms of true and false positive rates across demographics, and define the less restrictive equal opportunity. The authors extend these concepts to continuous features and define obliviousness to limit prediction information. The authors define when a predictor is derived and show the importance of a loss function. Using their fairness definitions, the authors produce a linear optimization problem solving for the best predictor and give a geometric interpretation using feasible regions and ROC curves with score thresholding and show to graphically obtain the predictors. The authors place their concepts in a Bayes classification setting and derive existence theorems of the predictors. The authors apply their concepts to FICO data using five threshold requirements and compare their predictor accuracy and utility. The authors summarize the strengths and limitations of their concepts.

Two Uses or Extensions

- The current equalized odds and equal opportunity set-up does not allow for continuous attributes such as A representing income. An extension to this is to allow A to be continuous from a set S. The equalized odds and equal opportunity conditions will correspondingly be $\Pr(\hat{Y} = 1|A = a, Y = y)$ for all $a \in S$. This would not necessarily produce a solution in linear optimization in Figure 1 so the optimal predictor $\hat{Y}$ would instead have to minimize the sum of errors $\int_{a \in S}(\Pr(\hat{Y} = 1|A = a, Y = y) - T)^2$ where $T = \mathbb{E}_{a \in S}[\Pr(\hat{Y} = 1|A = a, Y = y)]$.
- The equal opportunity definition is applicable to college admissions. A side effect of affirmative action policies in the U.S. is that demographics that are recipients of it tend to have higher dropout rates. This side effect is due to the policies designed to follow the demographic parity concept in the paper. If instead the policies followed the equal opportunity concept where the test for applicant X is "X will not drop out", then a predictor satisfying equal opportunity will allow admissions so that every value of the demographic variable A will have equal true positive rates, or non-dropout rates of accepted individuals. Thus, using the equal opportunity definition will remove the current dropout discrepancy side effect.

Paper: Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al., "Human-level control through deep reinforcement learning," Nature, 2015.

Summary: The authors explain a model they created that can play games and describe the type and inspiration of the neural network they used. The authors explain the objective the agent is optimizing, explain causes for its instability and the experience replay and correlation reduction techniques they used to reduce it. The authors argue for the efficiency of their algorithm by explaining the experience replay algorithm and loss function used. The authors explain the input images and training performance before describing the comparison algorithms, the agent's performance against them, and the necessity of all the agent's components. The authors explained the techniques used by the agent to represent images and properties of these representations, particularly in Space Invaders. The authors describe the types of games, the agent's performance on strategy games, and connection to the human brain of techniques to create the agent. The authors explain the image processing for the specific display, rationale and layers of their network, techniques used to train parameters, comparison logistics, and an overview and motivation of the algorithm, its objectives, and techniques. The authors provide figures showing the value objective over time for certain games, hyperparameter descriptions, and the agent's relative results on the tasks.

Two Uses or Extensions

- The agent can be adapted to perform automated car driving. Similar to the games in the paper, car driving has a limited number of actions that can be done at any given moment and a value objective can be created that can be affected by safety, obeying traffic rules, gas cost, and time. Unlike a human, the agent would be able to mount a camera that can see much further ahead the road so the agent is provided with more information than a human in contrast to the games in the paper, which suggests stronger relative performance in car driving than in games.

- The agent can be adapted to perform automated theorem proving. Rather than input being images, it would instead be tokens that decompose a mathematical statement to be proved using a system of logic such as predicate logic. Similar to the games in the paper, there are a limited number of "moves" (e.g. logical connectives, quantifiers) to transform a true proposition to another true proposition. The value objective would be based on how close the currently proved proposition is to theorem to be proved. As mentioned in the paper, the agent is able to invent strategies to play certain games, and this type of learning is helpful once the agent has stored enough experience to understand, similar to a human, the most plausible ways to begin proving a proposition.