Breast Cancer Fine-Needle Classification

Lea-Tereza Tenekedjieva (lten@stanford.edu)

Predicting

Purpose: develop an intelligent non-invasive diagnostic system for breast cancer classification.

- **Fine-needle aspiration (FNA)** - minimally invasive diagnostic procedure ⇒ alogirthm will allow for a reliable diagnosis without the need for an invasive surgical biopsy
- **Compare Three Models** - extracted features presented to three models to identify the most suitable model for effective tumor classification.
- **Baseline Model** - Logistic regression provided really high baseline, and network with tanh activation was able to improve on specificity.
- **Question** - Can we significantly improve performance by employing a more complex model?

Dataset and Features

Dataset consists of features extracted from a digitized image of FNA of breast lumps.

- Distribution: 357 benign, 212 malignant tumors.
- Each cell nucleus has 10 real-valued features:

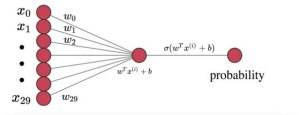
Radius *Texture* *Perimeter* *Symmetry*
Concavity *Concave points* *Compactness*
Symmetry *Fractal dimension* *Smoothness*

For each feature mean, standard error, largest of the features are computed ⇒ **30 features total**.

Logistic Regression

$$y^{(i)} = a^{(i)} = \sigma(w^T x^{(i)} + b)$$

$$J = -\frac{1}{m}\sum_{i=1}^{m}\Big(y^{(i)} log(a^{(i)}) + (1 - y^{(i)}) log(1 - a^{(i)})\Big)$$

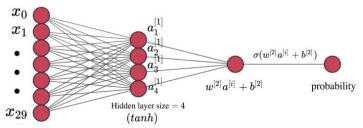


Two-layer Neural Net with Tanh Activation

$$a^{[1](i)} = tanh(w^{[1]} x^{(i)} + b^{[i]})$$

$$y^{(i)} = a^{[2](i)} = \sigma(w^{[2]} a^{[i](i)} + b^{[2]})$$

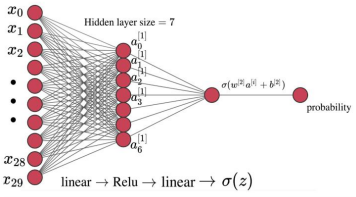
$$J = -\frac{1}{m}\sum_{i=1}^{m}\Big(y^{(i)} log(a^{[2](i)}) + (1 - y^{(i)}) log(1 - a^{[2](i)})\Big)$$



Two-layer Neural Net with RELU Activation

$$a^{[1](i)} = \text{RELU}(w^{[1]} x^{(i)} + b^{[i]})$$

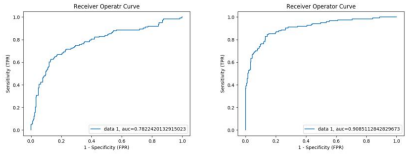
$$y^{(i)} = a^{[2](i)} = \sigma(w^{[2]} a^{[i](i)} + b^{[2]})$$



Results

	LogReg		tanh	RELU	
	Train	Dev	Dev	Train	Dev
Accuracy	0.8260	0.9534	0.936813	0.9478	0.3804
Sensitivity	0.9611	1.0	0.7619	0.8618	1.0
Specificity	0.7459	0.9259	0.9859	0.9917	0.19718

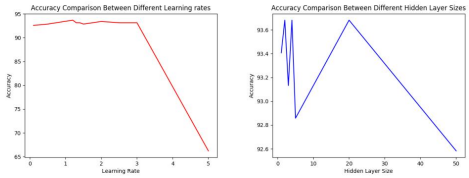
Receiver operating characteristic



tanh RELU

Results: Accuracies

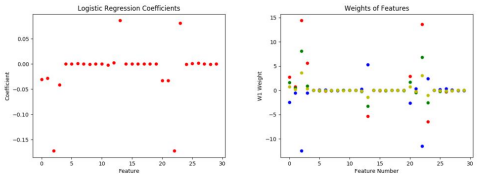
Accuracies of NN with tanh activation



Acc. vs Learning rate Acc. vs #hidden nodes

Results: Features

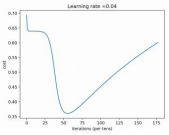
Weights of features



Logistic Regression NN with tanh activation

Both models are learning on the same features.
Most significant: **perimeter**, **area** and **radius**.
Other features: almost no learning at all.

Discussion and DeepNet tuning



Model doesn't generalize well to new data, likely **overfitting**. Attempted: L2 Regularization, tuning λ and α values, learning rate decay, number of iterations and early stopping.

Further work necessary to determine right balance of hyperparameters to make third model viable.

Logistic regression and NN with tanh activation network perform significantly better than a more complex model.

George, Y., Bassant Z. et al. (2012). *Breast Fine Needle Tumor Classification using Neural Networks.* IJCSI International Journal of Computer Science. 9. 247-256.
R.W.M. Giard and J.Hermans. *The value of aspiration cytologic examination of the breast.* A statistical review of the medical literature. Cancer. 69:2104-2110. 1992