Prediction of MHC Class I and II binding peptides incorporating bayesian transfer hierarchies

Ravikiran Janardhana

December 4, 2012

So far...

- Increased feature set for Elastic Net (baseline) by incorporating interaction features, i.e, Protein (a,b) at Pos (x,y). There are 80100 features now compared to 300 earlier
- Using Matlab's Elastic Net implementation (Lasso) and SVM (svmtrain and svmclassify), I classified the binding and non-binding peptides. The accuracies between the two methods are now comparable and it varies from 70-75%, the state of the art reports 80% for real data. The accuracy of Elastic Net earlier was mediocre at 53% and there is a drastic improvement with the addition of these features.
- **Question:** How many training samples and testing samples needs to be there in each set for an acceptable result?

Model:

The optimization problem for two related MHC-Class II alleles classifier is given by

$$\begin{aligned} \underset{\mathbf{w}^1,\mathbf{w}^2,\mathbf{w}}{\text{minimize}} \quad & \frac{1}{2}\left\|\mathbf{y}^1-\mathbf{x}^1\right\|_2^2+\frac{1}{2}\left\|\mathbf{y}^2-\mathbf{x}^2\right\|_2^2+ \\ & \lambda_1\left\|\mathbf{w}^1\right\|_1+\lambda_2\left\|\mathbf{w}^2\right\|_1+\alpha\left\|\mathbf{D}\mathbf{w}\right\|_1. \end{aligned}$$

where,

$$\mathbf{D}=\begin{bmatrix} 1000\,.....\,-1000 \\ 1000\,.....\,0\,-100 \\ 1000\,.....\,00\,-10 \\ 1000\,.....\,000\,-1 \end{bmatrix} \quad \mathbf{w}=\begin{bmatrix}\mathbf{w}^1\\ \mathbf{w}^2\end{bmatrix}.$$

We are going to introduce new variables $\mathbf{z}^1,\mathbf{z}^2,\mathbf{z}^3,\mathbf{z}^4,\mathbf{z}^5$ and reformulate the problem

$$\begin{aligned} \underset{\mathbf{w},\mathbf{z}^1,\mathbf{z}^2,\mathbf{z}^3,\mathbf{z}^4,\mathbf{z}^5}{\text{minimize}} \quad & \frac{1}{2}\left\|\mathbf{y}^1-\mathbf{z}^1\right\|_2^2+\frac{1}{2}\left\|\mathbf{y}^2-\mathbf{z}^2\right\|_2^2+ \\ & \lambda_1\left\|\mathbf{z}^3\right\|_1+\lambda_2\left\|\mathbf{z}^4\right\|_1+\alpha\left\|\mathbf{z}^5\right\|_1. \end{aligned}$$

Writing out the augmented lagrangian for the above problem,

$$\begin{aligned}\mathrm{AL}(\mathbf{w}, \mathbf{z}^0, \mathbf{z}^1, \mathbf{z}^2, \mathbf{z}^3, \mathbf{z}^4, \mathbf{z}^5, \mathbf{u}^1, \mathbf{u}^2, \mathbf{u}^3, \mathbf{u}^4, \mathbf{u}^5) \quad = \quad & \frac{1}{2}\left\|\mathbf{y}^1 - \mathbf{z}^1\right\|_2^2 + \frac{1}{2}\left\|\mathbf{y}^2 - \mathbf{z}^2\right\|_2^2 + \lambda_1\left\|\mathbf{z}^3\right\|_1 + \lambda_2\left\|\mathbf{z}^4\right\|_1 + \alpha\left\|\mathbf{z}^5\right\|_1 \\ & + \mathbf{u}^1(\mathbf{z}^1 - \mathbf{x}^1) + \mathbf{u}^2(\mathbf{z}^2 - \mathbf{x}^2) \\ & + \mathbf{u}^3(\mathbf{z}^3 - \mathbf{w}^1) + \mathbf{u}^4(\mathbf{z}^4 - \mathbf{w}^2) + \mathbf{u}^5(\mathbf{z}^5 - \mathbf{D}\mathbf{w}) \\ & + \frac{\rho}{2}\left\|\mathbf{z}^1 - \mathbf{x}^1\right\|_2^2 + \frac{\rho}{2}\left\|\mathbf{z}^2 - \mathbf{x}^2\right\|_2^2 \\ & + \frac{\rho}{2}\left\|\mathbf{z}^3 - \mathbf{w}^1\right\|_2^2 + \frac{\rho}{2}\left\|\mathbf{z}^4 - \mathbf{w}^2\right\|_2^2 + \frac{\rho}{2}\left\|\mathbf{z}^5 - \mathbf{D}\mathbf{w}\right\|_2^2\end{aligned}$$

The derived updates are as below :-

```
w1 = [I] \ [z3 + ( (1 / rho) * u3)]
w2 = [I] \ [z4 + ( (1 / rho) * u4)]
w  = [D] \ [z5 + ( (1 / rho) * u5)]
```

Question: Is the above correct or should I stack all 3 rows to derive update for 'w' as 'w' is made up of 'w1' and 'w2' ?

```
z1 = [eye(n1); sqrt(rho) * eye(n1)] \ [y1; (sqrt(rho) * x1) - ((1 / sqrt(rho)) * u1)]
z2 = [eye(n2); sqrt(rho) * eye(n2)] \ [y2; (sqrt(rho) * x2) - ((1 / sqrt(rho)) * u2)]

z3 = shrinkThreshold(w1 - 1/rho*u3, lambda1/rho)
z4 = shrinkThreshold(w2 - 1/rho*u4, lambda2/rho)
z5 = shrinkThreshold(D*w - 1/rho*u5, alpha/rho)

u1 = u1 + rho * (z1 - x1);
u2 = u2 + rho * (z2 - x2);
u3 = u3 + rho * (z3 - w1);
u4 = u4 + rho * (z4 - w2);
u5 = u5 + rho * (z5 - w);
```

Question: What is the order in which I should compute updates first ? w1, w2, w, z1-z5 ? Is this order correct ?

Question: How to initialize λ_1 , λ_2, ρ and α ?

Question: We say $\mathbf{z}^1 = \mathbf{x}^1$, but $\mathbf{x}^1$ is a matrix of feature vector rows (80100 in length), unable to think of how to initialize $\mathbf{z}^i$