4 Linear Models for Classification

Definition. ***Concepts***

1. ***Classification*** *Take input vector* $\mathbf{x}$ *andf assign it to one of* K *discrete classes* $\mathcal{C}_k$ *where* $k = 1, \cdots, K$
2. *Decision region, decision boundary*
3. *Linearly Separable*
4. *1-of-K encoding for target value* t
5. *Approaches*
   (a) *Discriminant function*
   (b) *Model conditional probability distribution* $p(\mathcal{C}_k|\mathbf{x})$ *either directly or via modeling class-conditional densities* $p(\mathbf{x}|\mathcal{C}_k)$ *with a given prior* $p(\mathcal{C}_k)$
6. *Generalized Linear Model and Activation Function* f

$$y(\mathbf{x}) = f(\mathbf{w}^T\mathbf{x} + w_0)$$

where decision surfaces are linear functions of $\mathbf{x}$ *(not linear to* $\mathbf{w}$*)*

4.1 Discriminant Functions

Definition. ***Points***

1. ***Two Class*** *For linear discriminant functions* $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$*, the weight vector* $\mathbf{w}$ *is orthogonal to the decision surface, so determines orientation of the decision surface.*
2. ***Multiclass***
   (a) *One-versus-Rest Classifier*
   (b) *One-versus-one Classifier*
   (c) *K-class Discriminant*

$$y_k(\mathbf{x}) = \mathbf{w}_k^T\mathbf{x} + w_{k0} \qquad \text{Assign } \mathbf{x} \text{ to } \mathcal{C}_k \text{ if } y_k(\mathbf{x}) > y_j(\mathbf{x}) \text{ for all } j \neq k$$

3. ***Least Squares*** *Given training set* $\{\underset{N\times(D+1)}{\tilde{\mathbf{X}}}, \underset{N\times K}{\mathbf{T}}\}$*, we have exact closed-form solution for discriminate function* $\mathbf{y}(\mathbf{x})$ *parameters* $\underset{(D+1)\times K}{\tilde{\mathbf{W}}}$

$$y(\mathbf{x}) = \tilde{\mathbf{W}}^T\tilde{\mathbf{X}} \qquad \tilde{\mathbf{W}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{T}$$

4. ***Fisher's Linear Discriminant*** *Given mean vector of classes*

$$\mathbf{m}_k = \frac{1}{N_1} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n$$

(a) *Maximizes separation of projected (to y) class, i.e. for 2 classes maximizes the mean of projected data m_k*

$$m_2 - m_1 = \mathbf{w}^T \mathbf{m}_2 - \mathbf{w}^T \mathbf{m}_1 \qquad \textit{subject to } \sum_i w_i^2 = 1$$

Optimal solution given by $\mathbf{w} \propto \mathbf{m}_2 - \mathbf{m}_1$

(b) *Maximizes separation between projected class mean while gives a small variance within each class to minimize overlap, within class variance given by*

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (\mathbf{w}^T \mathbf{x}_n - m_k)^2$$

Fisher criterion is the ratio of between class variance to within-class variance

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \qquad \mathbf{S}_W = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T$$

Optimal solution given by $\mathbf{w} \propto \mathbf{W}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$. Projecting data points into y space and choose a threshold y_0 such that we classify new point belonging to $\mathcal{C}_1$ if $y(\mathbf{x}) \geq y_0$ and belonging to $\mathcal{C}_2$ otherwise

4.2 Probabilitstic Generative Models

Definition. ***Points***

1. ***Logistic function*** $\sigma(a)$

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} = \frac{1}{1 + exp\{-a\}} = \sigma(a) \qquad a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

and logit function a, the inverse of logistic, representing log of ratio of posterior probabilities

$$a = \ln \left\{ \frac{\sigma}{1 - \sigma} \right\} = \ln \left\{ \frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})} \right\}$$

2. ***Softmax Function***,

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} = \frac{exp\{a_k\}}{\sum_j exp\{a_j\}} \qquad a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$$

a smoothed version of max function

3. ***Continuous Inputs*** *if class conditional density* $p(\mathbf{x}|\mathcal{C}_k)$ *is Gaussian with all classes sharing the same covariance matrix, then we can express posterior distribution as a logistic sigmoid acting on a linear function of* $\mathbf{x}$,

$$p(\mathcal{C}_k|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}+w_0) \qquad \mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \quad w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+\frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+\ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

The genearlized version for K *classes is given by*

$$a_k(\mathbf{x}) = \mathbf{w}_k^T\mathbf{x} + w_{k0} \qquad \mathbf{w}_k = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k \quad w_{k0} = -\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p(\mathcal{C}_k)$$

4. *Quadratic Discriminant*

5. ***MLE solution for parameter for class-conditional densities and prior*** *for Gaussian class-conditionals with 2 classes* ($t_n = 1$ *for* $\mathcal{C}_1$ *and* $t_n = 0$ *for* $\mathcal{C}_2$ *and prior* $p(\mathcal{C}_1) = \pi$ *and* $p(\mathcal{C}_2) = 1-\pi$*). Note for data point* x_n *from class* $\mathcal{C}_1$*, then* $p(\mathbf{x}_n, \mathcal{C}_1) = p(\mathcal{C}_1)p(\mathbf{x}_n|\mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$*, we have likelihood function*

$$p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} (\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}))^{t_n} ((1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}))^{1-t_n}$$

We find mle estimate for

$$\begin{aligned}
\pi &= \frac{N_1}{N_1+N_2} && \text{(fraction of points in each class)} \\
\boldsymbol{\mu}_1 &= \frac{1}{N_1}\sum_{n=1}^{N} t_n\mathbf{x}_n && \text{(mean of input vectors assigned to } \mathcal{C}_1\text{)} \\
\boldsymbol{\mu}_2 &= \frac{1}{N_2}\sum_{n=1}^{N}(1-t_n)\mathbf{x}_n && \text{(mean of input vectors assigned to } \mathcal{C}_2\text{)} \\
\boldsymbol{\Sigma} &= \frac{N_1}{N}\mathbf{S_1} + \frac{N_2}{N}\mathbf{S_2} && \text{(weighted average of covariance matrix for 2 classes)}
\end{aligned}$$

4.2 Probabilitstic Discriminative Models

Definition. ***Points***

1. ***Motivation*** *In discriminative approach, we are maximizing a likelihood function defined through conditional posterior distribution* $p(\mathcal{C}_k|\mathbf{x})$*. Needs fewer adaptive parameters to be determined*

2. ***Fixed basis function*** *nonlinear transform of inputs using* $\phi(\mathbf{x})$*, such that decision boundary is linear in feature space but nonlinear in input space*

3. ***Logistic Regression*** *Write posterior probability of each class as logistic sigmoid over a linear function of feature vector* $\boldsymbol{\phi}$*, such that the number of adjustable parameter is linear to the feature space (vs. quadratic for generating function approach)*

$$p(\mathcal{C}_1|\boldsymbol{\phi}) = y(\boldsymbol{\phi}) = \sigma(\mathbf{w}^T\boldsymbol{\phi}) \qquad p(\mathcal{C}_2|\boldsymbol{\phi}) = 1 - p(\mathcal{C}_1|\boldsymbol{\phi})$$

Let $\boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n)$ *and* $y_n = p(\mathcal{C}_1|\boldsymbol{\phi}_n) = \sigma(\mathbf{w}^T\boldsymbol{\phi}_n)$*, we have likelihood*

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}$$

$$\mathcal{E}(\mathbf{w})_{CE} = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}(t_n \ln y_n + (1-t_n)\ln(1-y_n)) \quad \rightarrow \quad \nabla_{\mathbf{w}}\mathcal{E}(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\mathbf{x}_n$$

No closed form solution for $\mathbf{w}$*, due to nonlinearity of logistic sigmoid. Error function is however convex.*

4. ***Multiclass Logistic Regression*** *By generative approach for multiclass classification, posterior distribution given by softmax transformation of linear function of feature variables*

$$p(\mathcal{C}_k|\boldsymbol{\phi}) = y_k(\boldsymbol{\phi}) = \frac{exp\{a_k\}}{\sum_j exp\{a_j\}} \qquad a_k = \mathbf{w}_k^T\boldsymbol{\phi} \text{ (activation)}$$

with likelihood

$$p(\mathbf{T}|\mathbf{w}_1, \cdots, \mathbf{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(\mathcal{C}_k|\boldsymbol{\phi}_n)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

$$\mathcal{E}_{CE}(\mathbf{w}_1, \cdots, \mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1, \cdots, \mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln y_{nk}$$