Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Piyush Bagad
Roll Number: 150487
Date: September 22, 2018

QUESTION 1

Consider a logistic regression model with gaussian prior.

$$p(y_n|\mathbf{x_n}, \mathbf{w}) = \frac{1}{1+\exp(-y_n\mathbf{w}^T\mathbf{x_n})}, \quad p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I})$$

For MAP estimation, we need to solve the following maximization problem:

$$\hat{\mathbf{w}}_{MAP} = \arg\min_w \mathcal{L}(\mathbf{w}) = \arg\min_w \left(-\sum_{n=1}^{N} \log(p(y_n|\mathbf{x_n}, \mathbf{w})) - log(p(\mathbf{w})) \right)$$

$$\log(p(y_n|\mathbf{x_n}, \mathbf{w})) = -\log(1+\exp(-y_n\mathbf{w}^T\mathbf{x_n}), \;\; \log(p(\mathbf{w})) = -(D/2)\log(2\pi)+\log(\lambda)-(\lambda/2)\mathbf{w}^T\mathbf{w}$$

$$\therefore \mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N} \log(1+\exp(-y_n\mathbf{w}^T\mathbf{x_n}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

Computing the partial derivative w.r.t $\mathbf{w}$, and equating to 0, we get

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \sum_{n=1}^{N} \frac{-y_n\mathbf{x}_n}{1+\exp(-y_n\mathbf{w}^T\mathbf{x_n})} + \lambda\mathbf{w} \Rightarrow \boxed{\mathbf{w} = \sum_{n=1}^{N} y_n\alpha_n\mathbf{x}_n}$$

$$\text{where} \;\; \alpha_n(\mathbf{w}) = \frac{1}{\lambda(1+\exp(-y_n\mathbf{w}^T\mathbf{x_n}))} = \frac{p(y_n|\mathbf{x}_n, \mathbf{w})}{\lambda}$$

Thus, the MAP estimate is trying to weigh examples based on the prediction probability of each given example. It will ignore the examples on which the predicted probability is low ($p(y_n|\mathbf{x_n}, \mathbf{w}) \to 0$) and consider only those examples that have this predicted probability high much like the support vectors in SVM. The hyperparamter λ will decide the amount of regularization as usual.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Piyush Bagad
Roll Number: 150487
Date: September 22, 2018

QUESTION 2

Consider a generative classification model for binary classification with the Naive Bayes assumption of feature independence. Let the class marginal and class conditional distributions be as follows:

$$p(y=1)=\pi, \quad p(\mathbf{x}|y=j)=\prod_{d=1}^{D}\text{Bernoulli}(x_d|\mu_{d,j})=\prod_{d=1}^{D}\mu_{d,j}^{x_d}(1-\mu_{d,j})^{1-x_d}, \quad j=0,1$$

For convenience, let us consider simplification of only $p(y=1|\mathbf{x})$.

$$p(y=1|\mathbf{x})=\frac{p(\mathbf{x}|y=1)p(y=1)}{p(\mathbf{x})}=\frac{p(\mathbf{x}|y=1)p(y=1)}{p(\mathbf{x}|y=1)p(y=1)+p(\mathbf{x}|y=0)p(y=0)}$$

$$\therefore p(y=1|\mathbf{x})=\frac{\pi\prod_{d=1}^{D}B(x_d,\mu_{d,1})}{\pi\prod_{d=1}^{D}B(x_d,\mu_{d,1})+(1-\pi)\prod_{d=1}^{D}B(x_d,\mu_{d,0})}; \quad \text{where } B \text{ denotes Bernoulli}$$

$$\therefore p(y=1|\mathbf{x})=\frac{1}{1+\frac{1-\pi}{\pi}\frac{\prod_{d=1}^{D}B(x_d,\mu_{d,1})}{\prod_{d=1}^{D}B(x_d,\mu_{d,0})}}=\frac{1}{1+\frac{1-\pi}{\pi}\prod_{d=1}^{D}(\frac{\mu_{d0}}{\mu_{d1}})^{x_d}(\frac{1-\mu_{d0}}{1-\mu_{d1}})^{1-x_d}}$$

Now, let $s_d=\frac{\mu_{d0}}{\mu_{d1}}$, $r_d=\frac{1-\mu_{d0}}{1-\mu_{d1}}$ and $C=\frac{1-\pi}{\pi}$. On further simplifying, we get

$$p(y=1|\mathbf{x})=\frac{1}{1+C\prod_{d=1}^{D}s_d^{x_d}r_d^{1-x_d}}=\frac{1}{1+\exp\left(\log(C)+\sum_{d=1}^{D}x_d\log(s_d)+(1-x_d)\log(r_d)\right)}$$

$$\therefore \boxed{p(y=1|\mathbf{x})=\frac{1}{1+\exp(\mathbf{w}^T\mathbf{x}+b)}}$$

$$\text{where } \mathbf{w}=[\log(s_1)-\log(r_1),...,\log(s_D)-\log(r_D)]^T \text{and } b=\log\left(C(\prod_{d=1}^{D}r_d)\right)$$

Thus, this corresponds to the discriminative logistic regression model for binary classification with labels $\in\{0,1\}$. The decision boundary is linear.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Piyush Bagad
Roll Number: 150487
Date: September 22, 2018

QUESTION 3

Consider the following constrained version of least squares linear regression:

$$\hat{\mathbf{w}} = \arg\min_{w} \sum_{n=1}^{N} \left(y_n - \mathbf{w}^T\mathbf{x}_n\right)^2 \quad \text{subject to} \quad \|\mathbf{w}\| \leq c$$

Note that the constraint $\|\mathbf{w}\| \leq c$ is equivalent to $\|\mathbf{w}\|^2 \leq c^2$ since $\|\mathbf{w}\| \geq 0$. The lagrangian function for the modified constrained problem can be stated as:

$$\mathcal{L}(\mathbf{w}, \alpha) = \sum_{n=1}^{N} \left(y_n - \mathbf{w}^T\mathbf{x}_n\right)^2 + \alpha(\|\mathbf{w}\|^2 - c^2)$$

Ignoring the constant and re-writing,

$$\mathcal{L}(\mathbf{w}, \alpha) = \sum_{n=1}^{N} \left(y_n - \mathbf{w}^T\mathbf{x}_n\right)^2 + \alpha\mathbf{w}^T\mathbf{w} = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \alpha\mathbf{w}^T\mathbf{w}$$

Solving using the dual variable technique (We can do it since both the objective and constraint are convex),

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}} = -2\mathbf{X}^T\mathbf{y} + 2(\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})\mathbf{w} = 0$$

$$\boxed{\mathbf{w} = \left(\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}}$$

$$\therefore \mathcal{L}_D(\alpha) = \mathbf{y}^T\mathbf{y} - \mathbf{w}^T\mathbf{X}^T y - \mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T(\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})\mathbf{w}$$

$$\therefore \mathcal{L}_D(\alpha) = \mathbf{y}^T\mathbf{y} - \mathbf{w}^T\mathbf{X}^T y - \mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\mathbf{X}^T\mathbf{y} = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w}$$

Ignoring the term $\mathbf{y}^T\mathbf{y}$, we get

$$\mathcal{L}_D(\alpha) = -\mathbf{y}^T\mathbf{X}\mathbf{w} = -(\mathbf{X}^T\mathbf{y})^T\left(\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y} = -\mathbf{Z}^T\mathbf{M}_\alpha^{-1}\mathbf{Z}$$

$$\frac{\partial\mathcal{L}}{\partial\alpha} = -\frac{\partial[\mathbf{Z}^T\mathbf{M}_\alpha^{-1}\mathbf{Z}]}{\partial\mathbf{M}_\alpha}\frac{\partial\mathbf{M}_\alpha}{\partial\alpha} = (\mathbf{M}_\alpha^{-1}\mathbf{Z}\mathbf{Z}^T\mathbf{M}_\alpha^{-1})^{-1}\mathbf{I}$$

$$\therefore \frac{\partial\mathcal{L}}{\partial\alpha} = (\mathbf{M}_\alpha^{-1}\mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}\mathbf{M}_\alpha^{-1})^{-1}$$

Note that $\mathbf{M}_\alpha^T = \mathbf{M}_\alpha$ and thus we have

$$\therefore \frac{\partial\mathcal{L}}{\partial\alpha} = ((\mathbf{M}_\alpha^{-1}\mathbf{X}^T\mathbf{y})(\mathbf{M}_\alpha^{-1}\mathbf{X}^T\mathbf{y})^T)^{-1}$$

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Piyush Bagad
Roll Number: 150487
Date: September 22, 2018

QUESTION 4

Consider N training examples $\{x_n, y_n\}_{n=1}^N$ where $x_n \in \mathbb{R}^D, y_n \in \mathbb{R}^K$. Consider the softmax regression model for multi-class classification:

$$p(y_n = k|x_n, W) = \mu_{nk} = \frac{\exp(w_k^T x_n)}{\sum_{l=1}^K \exp(w_l^T x_n)}$$

The MLE objective can be stated as: $\hat{W}_{MLE} = \arg\min_W -NLL(W)$. Let us work for each column w_k of W:

$$\frac{-\partial NLL(W)}{\partial w_k} = -\sum_{n=1}^N \frac{\partial \log(p(y_n = k|x_n, W))}{\partial w_k}$$

$$\log(p(y_n = k|x_n, W)) = w_k^T x_n - \log\left(\sum_{l=1}^K \exp(w_l^T x_n)\right)$$

$$\therefore \frac{\partial \log(p(y_n = k|x_n, W))}{\partial w_k} = x_n - \mu_{nk} x_n$$

$$\therefore \frac{-\partial NLL(W)}{\partial w_k} = \sum_{n=1}^N x_n(\mu_{nk} - 1) = X^T \text{Diagonal}([\mu_{1k} - 1, \mu_{2k} - 1, .., \mu_{Nk} - 1])$$

Thus, no closed form solution for w_k can be computed from this equation. We can write the gradient descent update for w_k with $\eta = 1$ as follows:

$$\boxed{w_k^{(t+1)} = w_k^{(t)} - \sum_{n=1}^N x_n \left(\mu_{nk}^{(t)} - 1\right)}$$

Similarly, the stochastic gradient descent update will turn out to be:

$$\boxed{w_k^{(t+1)} = w_k^{(t)} - x_n(\mu_{nk}^{(t)} - 1)}$$

Consider the case of 'hard' assignments i.e. let $k = \arg\max_k \mu_{nk}$ and $\mu_{nk'} = 0, \forall k' \neq k$. The SGD update will simple reduce to

$$w_k^{(t+1)} = w_k^{(t)} - x_n(\mu_{nk}^{(t)} - 1) = \begin{cases} w_k^{(t)} & \text{for } p(y_n = k|x_n, W) = 1 \\ w_k^{(t)} + x_n & \text{for } p(y_n = k|x_n, W) = 0 \end{cases}$$

Thus, we simply do NOT update the weight vector w_k if we encounter an example (x_n, y_n) such that the current prediction is correct i.e. whose label is k. Else, we make the weight vector w_k move in the direction of x_n. The SGD algorithm sketch for this case will be as follows:

Algorithm 1 Multi-class Stochastic Gradient Descent

0: Choose initial $w_k = w_k^{(0)}, \forall k \in \{1, 2, ..K\}$.
0: **for** $k \in \{1, 2, ..K\}$ **do**
0: # Keeping w_l, for $l \neq k$ fixed, update w_k
0: **for** s iterations until convergence **do**
0: **if** $p(y_n|x_n, W^{(t)}) = 1$ **then**
0: $w_k^{(t+1)} = w_k^{(t)}$
0: **else**
0: $w_k^{(t+1)} = w_k^{(t)} + x_n$
0: **end if**
0: **end for**
0: **end for**=0

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Piyush Bagad
Roll Number: 150487
Date: September 22, 2018

QUESTION 5

Let C_x and C_y be the convex hulls corresponding to points $X = \{\mathbf{x}_i\}_{i=1}^N$ and $Y = \{\mathbf{y}_i\}_{i=1}^N$ respectively.

$$C_X := \{x = \sum_{n=1}^{N} \alpha_n x_n \ : \ \alpha_i \geq 0, \sum_i \alpha_i = 1\}$$
$$C_Y := \{x = \sum_{n=1}^{N} \beta_n y_n \ : \ \beta_i \geq 0, \sum_i \beta_i = 1\}$$

Definition 2.1. The sets of points X and Y are said to be linearly seperable if $\exists$ a line (hyperplane in high dimensional space) $L := \{\mathbf{x} \in \mathbb{R}^D : \ m^T\mathbf{x} + b = 0\}$ such that $m^T\mathbf{x}_i + b \geq 0$, and $m^T\mathbf{y}_i + b < 0 \ \forall \ i = 1, 2, ..N$.

We have to show that X and Y are linearly separable *iff* $C_X \cap C_Y = \phi$

($\Rightarrow$) Assume that X and Y are linearly separable. Suppose for contradiction that $C_X \cap C_Y \neq \phi$. This implies that $\exists z \in C_X \cap C_Y$. Thus, by definition,

$$z = \sum_{n=1}^{N} \alpha_n x_n \ = \sum_{n=1}^{N} \beta_n y_n$$

where $\alpha_i \geq 0, \sum_i \alpha_i = 1, \beta_i \geq 0, \sum_i \beta_i = 1$. Since, X and Y are linearly separable, there exists a line L satisfying $m^T\mathbf{x}_i + b \geq 0$, and $m^T\mathbf{y}_i + b < 0 \ \forall \ i = 1, 2, ..N$. Now, since $\alpha_i \geq 0, \forall i$,

$$\alpha_i m^T\mathbf{x}_i + \alpha_i b \geq 0, \forall i = 1, 2, .., N$$

$$\therefore \sum_{n=1}^{N} \alpha_n m^T\mathbf{x}_n + \sum_{n=1}^{N} \alpha_i b = \sum_{n=1}^{N} \alpha_n m^T\mathbf{x}_n + b.1 = m^T z + b \geq 0 \tag{1}$$

Similarly, since $\beta_i m^T\mathbf{y}_i + \beta_i b < 0 \ \forall \ i = 1, 2, ..N$,

$$\therefore \sum_{n=1}^{N} \beta_n m^T\mathbf{x}_n + \sum_{n=1}^{N} \beta_n b = \sum_{n=1}^{N} \beta_n m^T\mathbf{x}_n + b.1 = m^T z + b < 0 \tag{2}$$

The equations 1 and 2 present a contradiction. Thus, we have $C_X \cap C_Y = \phi$.

($\Leftarrow$) To show the converse, assume that $C_X \cap C_Y = \phi$. To show that X and Y are linearly separable i.e. we have to find L satisfying definition 2.1. Define

$$d := \min_{x \in C_X, y \in C_Y} \|x - y\|^2$$

Note, since C_X and C_Y are closed and bounded, the quantity d exists and suppose it achieves the minima at $x_o \in C_X, y_o \in C_Y$ .Also $d > 0$ since $C_X \cap C_Y = \phi$. Let L be the perpendicular bisector of the line joining x_o and y_o. It is trivial to observe that L is a linear separator of X and Y. Hence, we are done.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Piyush Bagad
Roll Number: 150487
Date: September 22, 2018

QUESTION 6

The original hard-margin SVM optimization problem can be stated as

$$\arg\min_{w,b} \frac{\|w\|^2}{2}$$
$$\text{subject to} \quad y_n(w^T x_n + b) \geq 1, \quad \forall\ n = 1, 2, .., N$$

The modified version of SVM involves changing the inequalities as $y_n(w^T x_n + b) \geq m, \quad \forall\ n = 1, 2, .., N$. Let $\alpha = [\alpha_1, \alpha_2, .., \alpha_N]$ be the lagrange variables. The lagrangian can be stated as

$$\mathcal{L}(w, b, \alpha) = \frac{\|w\|^2}{2} + \sum_{n=1}^{N} \alpha_n (m - y_n(w^T x_n + b))$$

Using the dual formulation to solve for the constrained optimization problem, we have

$$\frac{\partial \mathcal{L}(w,\alpha)}{\partial w} = 0 \Rightarrow \boxed{w = \sum_{n=1}^{N} \alpha_n y_n x_n}, \quad \frac{\partial \mathcal{L}(w,\alpha)}{\partial b} = 0 \Rightarrow \boxed{\sum_{n=1}^{N} \alpha_n y_n = 0}$$

Now, Substituting $w = \sum_{n=1}^{N} \alpha_n y_n x_n$ in Lagrangian, we get the dual problem as

$$\mathcal{L}_D(\alpha) = w^T w + \sum_{n=1}^{N} (\alpha_n m - y_n b) - \sum_{n=1}^{N} y_n w^T x_n = \sum_{n=1}^{N} \alpha_n m - \frac{1}{2} \sum_{n,l=1}^{N} \alpha_l \alpha_n y_l y_n (x_l^T x_n)$$

Now, the objective can be stated in a compact form as

$$\max_{\alpha \geq 0} \ m(\alpha^T \mathbf{1}) - \frac{1}{2} \alpha^T G \alpha$$

where G is an $N \times N$ matrix with $G_{ln} = y_l y_n x_l^T x_n$, and $\mathbf{1}$ is a vector of 1s. Note that m simply turns out to be a mutiplicative constant and thus does not affect the solution of the optimization problem. Thus, the solution to modified SVM effectively remains the same as that of original one.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Piyush Bagad
Roll Number: 150487
Date: September 22, 2018

QUESTION
7

Part 1: Dataset - `binclass.txt`

1a: Modeling positive and negative classes with different covariances

In this case, we assume that the positively labeled examples are sampled from the gaussian $\mathcal{N}(\mathbf{x}|\mu_+, \sigma_+^2\mathbf{I})$ and the negatively labeled examples from $\mathcal{N}(\mathbf{x}|\mu_-, \sigma_-^2\mathbf{I})$. Refer to figure 1. Notice the non linearlity of the decision boundary.

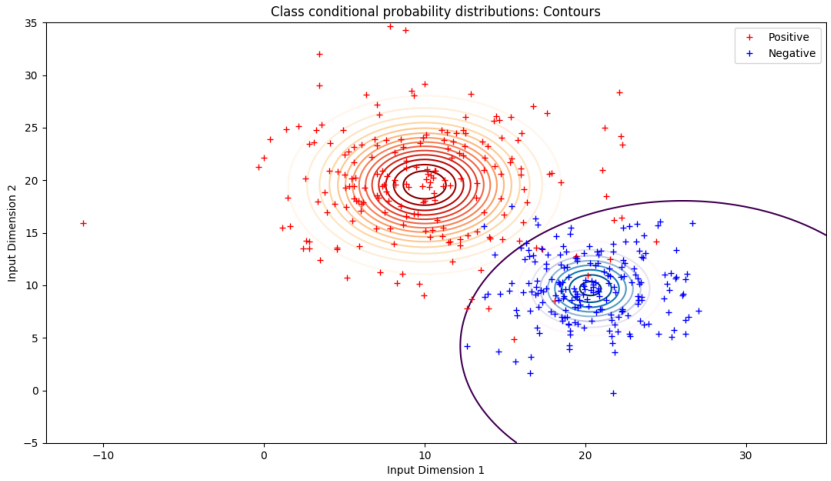

Figure 1: Decision boundary learned by generative classification with gaussian class conditional distributions having different covariances.

1b: Modeling positive and negative classes with same covariances

In this case, we assume that the positively labeled examples are sampled from the gaussian $\mathcal{N}(\mathbf{x}|\mu_+, \sigma^2\mathbf{I})$ and the negatively labeled examples from $\mathcal{N}(\mathbf{x}|\mu_-, \sigma^2\mathbf{I})$. Note that the covariance matrices for each of the two distributions is the same. Refer to figure 2. Notice the linearlity of the decision boundary.

1c: Linear SVM Decision Boundary

I have used `linearSVC` library from the `sklearn.svm` package (link). Refer to figure 3.

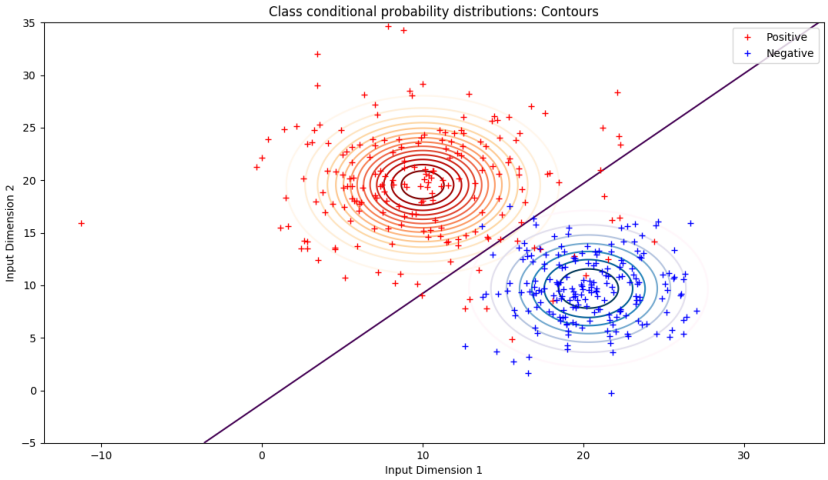

Figure 2: (Linear) Decision boundary learned by generative classification with gaussian class conditional distributions having same covariances.

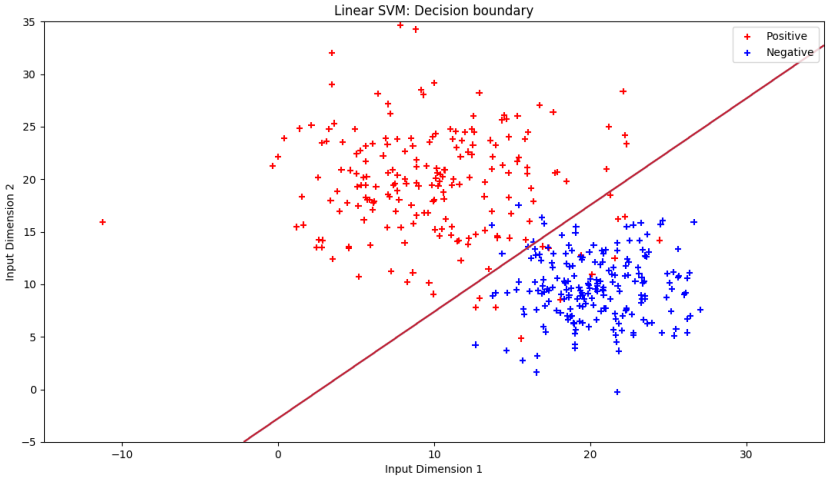

Figure 3: Decision boundary learned by Linear SVM using `linearSVC` from `sklearn.svm`.

Part 2: Dataset - `binclassv2.txt`

The experiments corresponding to various subsections for Part 1 have been repeated on the `binclassv2.txt` dataset and the results are noted below.

2a: Modeling positive and negative classes with different covariances

Refer to figure 4. Notice the non linearlity of the decision boundary.

2b: Modeling positive and negative classes with same covariances

Refer to figure 5. Notice the linearlity of the decision boundary.

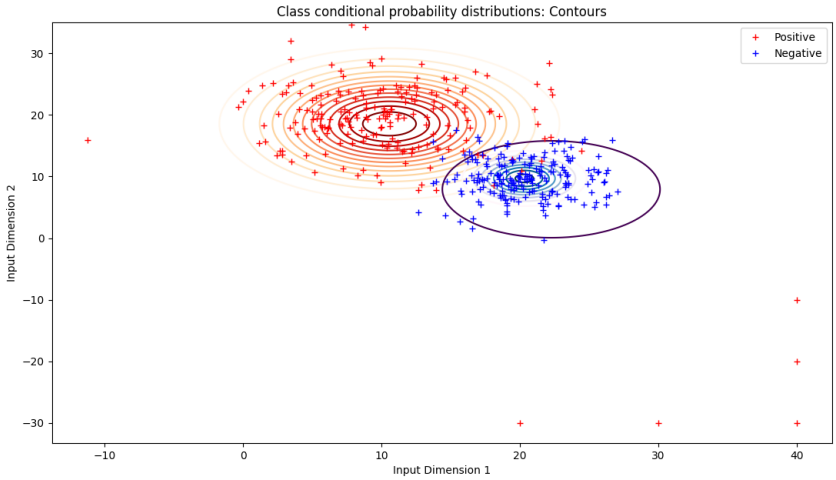

Figure 4: Decision boundary learned by generative classification with gaussian class conditional distributions having different covariances.

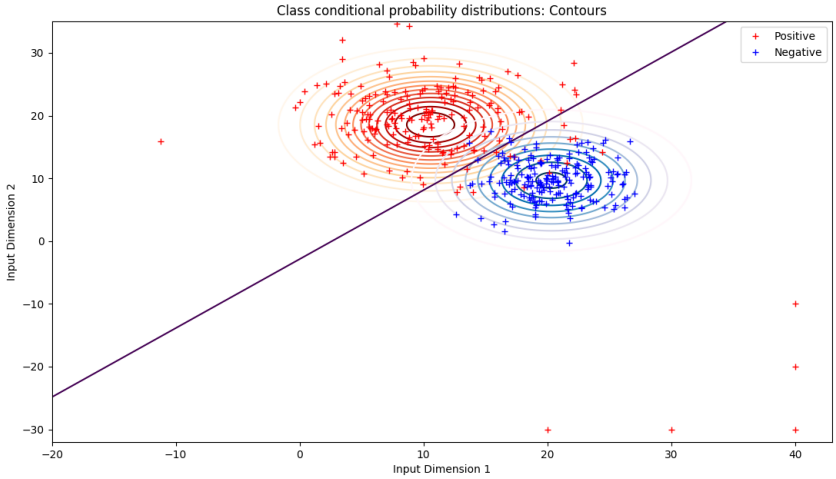

Figure 5: (Linear) Decision boundary learned by generative classification with gaussian class conditional distributions having same covariances.

2c: Linear SVM Decision Boundary

Refer to figure 6.

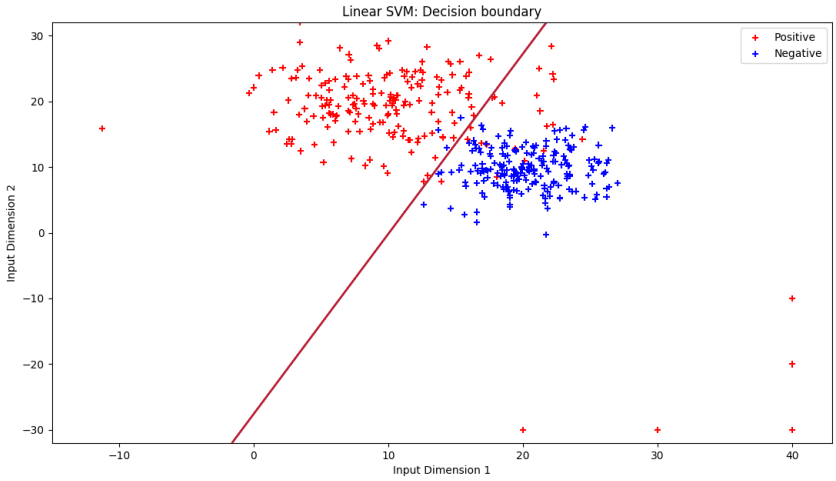

Figure 6: Decision boundary learned by Linear SVM using `linearSVC` from `sklearn.svm`.