CS573 DM HW1

Question 1

(1) Unbalanced data - very few instances of positive class compared to negative class.

(2) Re-weighting the objective function (similar to sampling bias) Cost sensitive classification - higher penalty for misclassifying one class than another. (Eg : in case of SVM, have two different slack penalty C_1 and C_2 instead of single C.)

(3) With increasing C, the accuracy of positive class increases (noticeable increase). However, the effect of increasing C on negative class is negligible (a very negligible decrease initially).

MAP estimate = $\text{argmax}_\beta \, P(\beta) \prod_i P(y_i|x_i, \beta)$

Taking log, MAP = $\text{argmax}_\beta \, log(P(\beta)) + \sum_i log(P(y_i|x_i, \beta))$
Since β follows a normal distribution, similar to the derivation in slides,

MAP = $\text{argmax}_\beta \sum_i log(P(y_i|x_i, \beta)) - \frac{1}{2\sigma^2} \sum_j \beta_j$
Comparing it with the LR formulation with L2 regularization (as mentioned in sklearn)

$C \propto \sigma^2$

From the MAP expression, for negative class with significantly more data points, the likelihood term will dominate more than the second prior probability term. However, for positive class, since there are fewer data points available, the prior has a much higher effect on the overall posterior probability. We already established the relation with regularization (C) and prior distribution (σ^2). Thus in positive class, since the prior has more effect on posterior than negative class, regularization affects the positive class much more than negative class.
(4) Consider π_p and π_n,
where π_p = number of instances of positive class/total number of data points
π_n = number of instances of negative class/total number of data points

π_p is significantly less than π_n due to data imbalance.
Likelihood can be rewritten as (giving more emphasis to positive class)

$$P(t|w) = \prod_c y(\phi_c)^{\frac{t_c}{\pi_p}} (1 - y(\phi_c))^{\frac{1-t_c}{\pi_n}}$$

Thus log likelihood,

$$log(P(t|w) = \sum_c \frac{t_c}{\pi_p} log(y(\phi_c)) + \frac{1 - t_c}{\pi_n} log(1 - y(\phi_c))$$

(5) Weights are chosen as inverse of the probabilities of each class (π_p and π_n) . Code attached separately.

(6)
Original SVM constrained optimization objective,

$$L(w, b, a) = \frac{1}{2}||w||^2 - \sum_{n=1}^{N} a_n\{t_n(w^T\phi(x_n) + b) - 1\}$$

We have to minimize this objective. Now if you look the second term, it incorporates the classification constraints. If a point is misclassified, the second term becomes negative, thus increasing the objective's value. This effect is similar for both positive and negative classes (except for Lagrange multipliers). We want a bigger penalty when we misclassifying positive class than negative class. The objective can be modified to the following objective to obtain this.

$$L(w, b, a) = \frac{1}{2}||w||^2 - \sum_{n=1}^{N} a_n\{\frac{t_n}{\pi_n}(w^T\phi(x_n) + b) - 1\}$$

where π_n is the probability of the class of n-th instance. (π_p and π_n defined above)

Question 2

(1) $P(\mathbf{y}|X, \boldsymbol{\beta}) = N(X\boldsymbol{\beta}, \lambda\mathbf{I})$.

(2)

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, P(\boldsymbol{\beta}|\mathbf{y}, X) \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, \frac{P(\mathbf{y}|X, \boldsymbol{\beta})P(\boldsymbol{\beta})}{\sum_{\boldsymbol{\beta}} P(\mathbf{y}|X, \boldsymbol{\beta})P(\boldsymbol{\beta})} \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmax}})P(\mathbf{y}|X, \boldsymbol{\beta}P(\boldsymbol{\beta}) \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, \frac{\exp\left(-\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})^{\mathrm{T}}\lambda\mathbf{I}^{-1}(\mathbf{y} - X\boldsymbol{\beta})\right)}{\sqrt{(2\pi)^k|\lambda\mathbf{I}|}} \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\beta})^{\mathrm{T}}\sigma\mathbf{I}^{-1}(\boldsymbol{\beta})\right)}{\sqrt{(2\pi)^k|\sigma\mathbf{I}|}} \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \quad \log\left(\frac{\exp\left(-\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\lambda}\mathbf{I}^{-1}(\mathbf{y} - X\boldsymbol{\beta})\right)}{\sqrt{(2\pi)^k|\lambda\mathbf{I}|}} \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\beta})^{\mathrm{T}}\sigma\mathbf{I}^{-1}(\boldsymbol{\beta})\right)}{\sqrt{(2\pi)^k|\sigma\mathbf{I}|}}\right) \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmax}} -\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})^{\mathrm{T}}\lambda^{-1}(\mathbf{y} - X\boldsymbol{\beta}) - \frac{1}{2}(\boldsymbol{\beta})^{\mathrm{T}}\sigma^{-1}(\boldsymbol{\beta}) \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmax}} -\frac{1}{\lambda}(\mathbf{y} - X\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - X\boldsymbol{\beta}) - \frac{1}{\sigma}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{\lambda}(\mathbf{y} - X\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} + X\boldsymbol{\beta}) - \frac{1}{\sigma}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}
\end{aligned}$$

$$= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{\lambda}||\mathbf{y} - X\boldsymbol{\beta})||_2 - \frac{1}{\sigma}||\boldsymbol{\beta}||_2$$

Question 3 : Decision Boundary and Classification

Consider a dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^n$ for a binary classification task with labels y={+1,-1}.
Assume that $\mathbf{x} = \mathcal{R}^p$. Let us define the mean of two classes as

$$\boldsymbol{\mu}_+ = \frac{1}{n_+} \sum_{y_i=+1} \mathbf{x}_i$$

$$\boldsymbol{\mu}_- = \frac{1}{n_-} \sum_{y_i=-1} \mathbf{x}_i$$

where $n_+ = \sum_{i=1}^n I_{\{y_i=+1\}}$ and $n_- = \sum_{i=1}^n I_{\{y_i=-1\}}$.

For a new data point **x** we check the **Euclidean distance** of the new point from $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ and assign the label of the class for which this distance is smaller.

Find the decision function. (Hint: It is of the form $\mathbf{w}^T\mathbf{x}$ + b. Express **w** and b in terms of $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$)
Solution:
We can write the decision boundary between two classes by the function:
f(**x**) = $||\boldsymbol{\mu}_- - \mathbf{x}||^2 - ||\boldsymbol{\mu}_+ - \mathbf{x}||^2$
$=||\boldsymbol{\mu}_-||^2 + ||\boldsymbol{x}||^2 - 2\langle\boldsymbol{\mu}_-,\boldsymbol{x}\rangle - ||\boldsymbol{\mu}_+||^2 - ||\boldsymbol{x}||^2 + 2\langle\boldsymbol{\mu}_+,\boldsymbol{x}\rangle$
$= ||\boldsymbol{\mu}_-||^2 - ||\boldsymbol{\mu}_+||^2 + 2\langle\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-,\boldsymbol{x}\rangle$
Comparing this with **w** and b we get $\mathbf{w}=2(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)$ and $b=||\boldsymbol{\mu}_-||^2 - ||\boldsymbol{\mu}_+||^2$
Alternatively, depending upon the function definition of f(x), $\mathbf{w}=2(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)$ and
$b=||\boldsymbol{\mu}_+||^2 - ||\boldsymbol{\mu}_-||^2$ is also an acceptable answer.