Problem 1. Support vector machines

Part a

confusion matrix for logistic regression test data:

```
 [[142  20]
 [ 34  58]]
```

Train:

```
[[299  39]
 [ 77  99]]
```

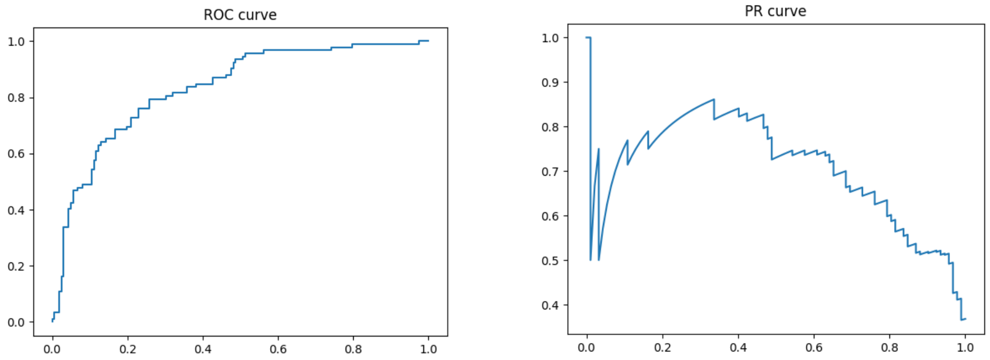
for linear kernel SVM test data:

```
[[132  28]
[ 22  49]]
```

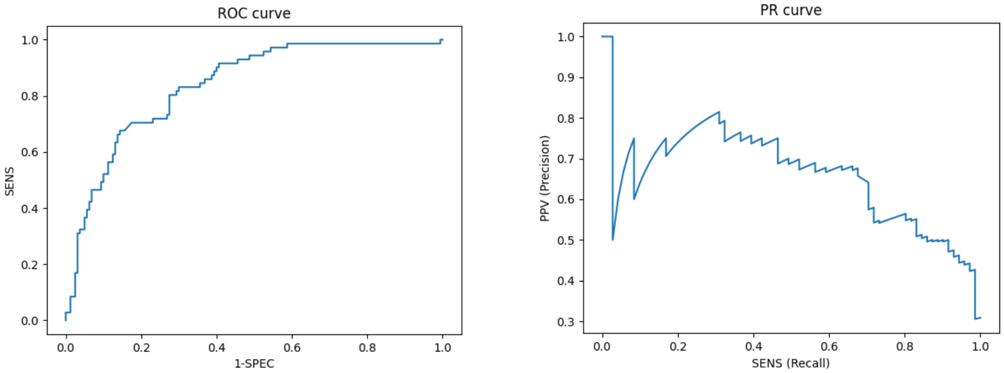
```
[[296  44]
[ 81 116]]
```

	Misclassification error	SENS	SPEC	PPV	NPV
Logistic regression test data	0.2125984251968504	0.6304347826086957	0.8765432098765432	0.7435897435897436	0.8068181818181818
SVM (linear kernel) test data	0.21645021645021645	0.6901408450704225	0.825	0.6363636363636364	0.8571428571428571
Logistic train data	0.22568093385214008	0.5625	0.8846153846153846	0.717391304347826	0.7952127659574468
SVM linear train data	0.23277467411545624	0.5888324873096447	0.8705882352941177	0.725	0.7851458885941645

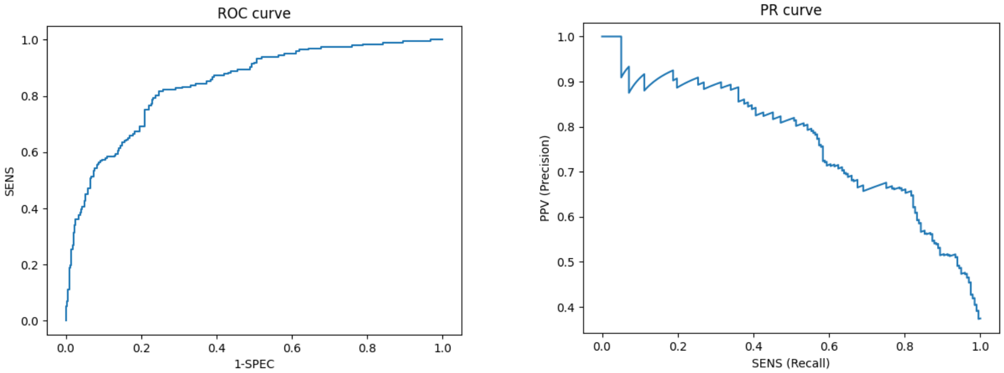
AUROC and AUPRC curves and scores:



Logistic regression model (test)



SVM (linear kernel) model (test)



SVM (linear kernel) model (train)

	AUROC	AUPRC	Mean accuracy
Logistic regression model test data	0.83	0.71	0.79
SVM (linear model) model test data	0.84	0.66	0.78
SVM linear train data	0.84	0.77	0.77

Based on the metrics above, the logistic regression model has a smaller misclassification error than the SVM model with linear kernel. Although the SVM's AUROC score is a little higher than the logistic regression mode, its AUPRC score is lower, and the difference in AUPRC scores is larger than the AUROC difference. The logistic regression model also has a higher mean accuracy. Thus, the logistic regression model is a little better than the SVM model with linear kernel on our test data. Also, based on the metrics on the train data, there is no worry of overfitting since the discrepancy between training and test data is small.

Part b

- **Second degree polynomial kernel**

Confusion matrix on test data:

```
[[142  18]
 [ 48  23]]
```

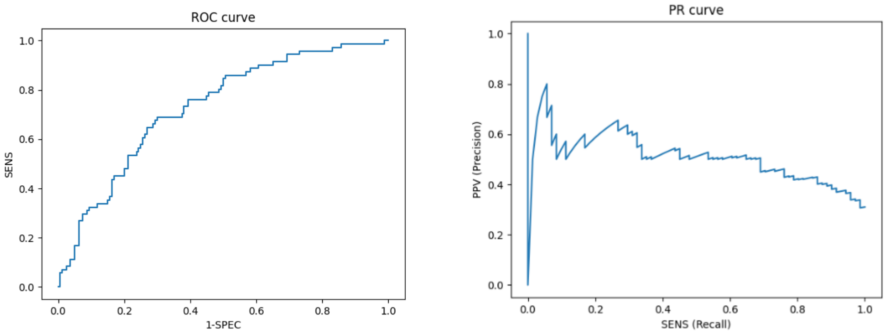
On train data:

```
[[325  15]
 [132  65]]
```

	Misclassification error	SENS	SPEC	PPV	NPV
SVM polynomial (2nd degree) kernel on test	0.2857142857142857	0.323943661971831	0.8875	0.5609756097560976	0.7473684210526316
On train	0.2737430167597765	0.3299492385786802	0.9558823529411765	0.8125	0.7111597374179431

AUROC score on test: 0.73
AUPRC score on test: 0.52
Mean accuracy on test: 0.71

AUROC on train: 0.75
AUPRC on train: 0.67
Mean accuracy on train: 0.73



SVM (poly kernel 2nd degree) on test

	Misclassification errors	AUROC	AUPRC	Mean accuracy
Logistic regression model on test	0.2125984251968504	0.83	0.71	0.79

SVM (linear model) model on test	0.21645021645021645	0.84	0.66	0.78
SVM (2nd degree poly kernel) model on test	0.2857142857142857	0.73	0.52	0.71

Compared with the logistic regression model and SVM model with linear kernel, the SVM model with 2nd degree polynomial kernel has higher misclassification error and smaller mean accuracy, AUROC, and AUPRC values. So 2nd degree polynomial kernel SVM model performs worse on the test data than the previous 2 models. Compared with the model applied on the train data, there is no significance to indicate overfitting because we don’t have large discrepancies between test and train errors.

- **RBF kernel**

  Confusion matrix on test:

  ```
  [[133  27]
   [ 24  47]]
  ```

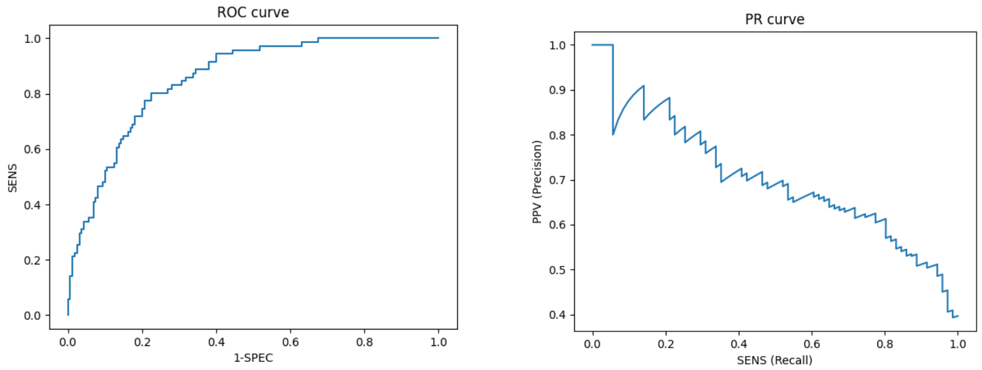
  On train:

  ```
  [[311  29]
   [ 69 128]]
  ```

	Misclassification error	SENS	SPEC	PPV	NPV
SVM RBF kernel on test	0.22077922077922077	0.6619718309859155	0.83125	0.6351351351351351	0.8471337579617835
On train	0.18249533445065177	0.649746192893401	0.9147058823529411	0.8152866242038217	0.8184210526315789

AUROC score on test: 0.85
AUPRC score on test: 0.70
Mean accuracy on test: 0.78

AUROC score on train: 0.90
AUPRC score on train: 0.84
Mean accuracy on train: 0.82



SVM (RBF kernel) on test

Summary for current models on test data:

	Misclassification errors	AUROC	AUPRC	Mean accuracy
Logistic regression model	0.2125984251968504	0.83	0.71	0.79
SVM (linear model) model	0.21645021645021645	0.84	0.66	0.78
SVM (2nd degree poly kernel) model	0.2857142857142857	0.73	0.52	0.71
SVM (RBF kernel) model	0.22077922077922077	0.85	0.70	0.78

The SVM model with RBF kernel has similar performance as the logistic regression one and the SVM linear kernel one. The model with RBF kernel's misclassification error, mean accuracy, and AUPRC values are a little worse than the logistic regression one but are close. The RBF kernel model's AUROC value is 0.85, higher than logistic regression and linear kernel SVM models. Compared with SVM model with 2nd degree polynomial kernel, the RBF kernel one's performance is much better. Compared with the train data error, although the error discrepancy is a little larger, it's still small enough to say that there is no overfitting problem.

- **Sigmoid kernel**

Confusion matrix on test:

```
[[118  42]
 [ 27  44]]
```

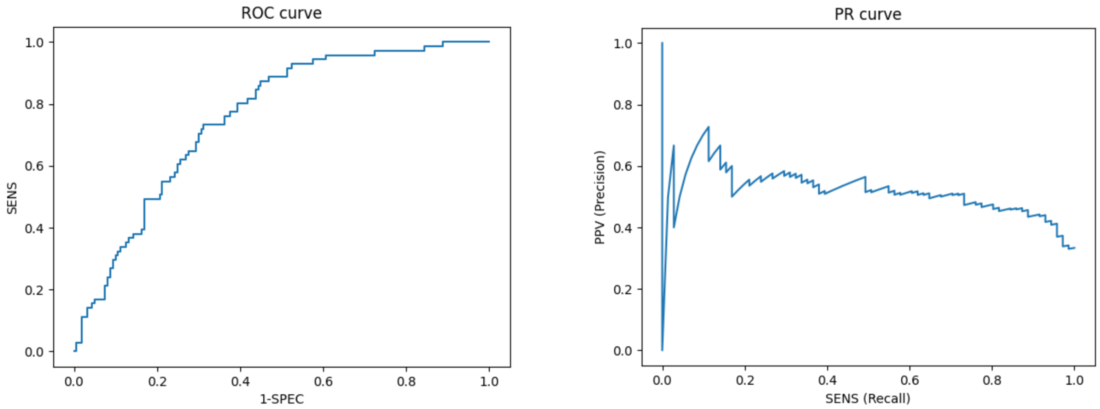
On train:

```
[[261  79]
 [ 88 109]]
```

	Misclassification error	SENS	SPEC	PPV	NPV
SVM sigmoid kernel on test	0.2987012987012987	0.6197183098591549	0.7375	0.5116279069767442	0.8137931034482758
On train	0.31098696461824954	0.5532994923857868	0.7676470588235295	0.5797872340425532	0.7478510028653295

AUROC score on test: 0.76
AUPRC score on test: 0.53
Mean accuracy on test: 0.70

AUROC score on train: 0.73
AUPRC score on train: 0.58
Mean accuracy on train: 0.69



SVM (sigmoid kernel) on test

On test data:

	Misclassification errors	AUROC	AUPRC	Mean accuracy
Logistic regression model	0.2125984251968504	0.83	0.71	0.79
SVM (linear model) model	0.21645021645021645	0.84	0.66	0.78
SVM (2nd degree poly kernel) model	0.2857142857142857	0.73	0.52	0.71
SVM (RBF kernel) model	0.22077922077922077	0.85	0.70	0.78
SVM (sigmoid kernel) model	0.2987012987012987	0.76	0.53	0.70

The SVM model with sigmoid kernel has the highest misclassification error. Its AUROC and AUPRC values are lower than logistic regression, the linear kernel, and the RBF kernel models, but slightly higher than the second-degree poly kernel model. Its mean accuracy is slightly smaller than the second-degree poly kernel one. Both the logistic regression model and SVM model with linear kernel perform better than the sigmoid kernel one. No worry of overfitting.

Part 2. Multilayer perceptron

Part a

Confusion matrix on test:

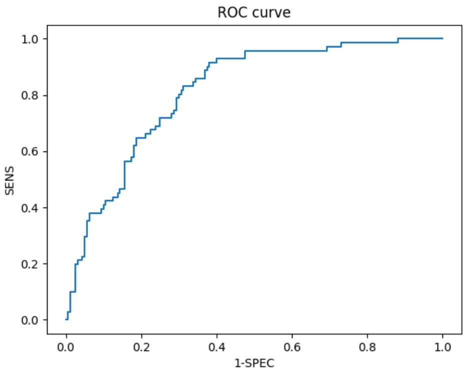
```
[[143  17]
 [ 42  29]]
```

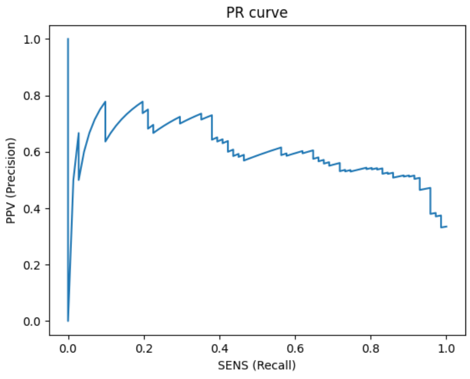
On train:

```
[[323  17]
 [115  82]]
```

	Misclassification error	SENS	SPEC	PPV	NPV
MLP (logistic) on test	0.2554112554112554	0.4084507042253521	0.89375	0.6304347826086957	0.772972972972973
On train	0.24581005586592178	0.41624365482233505	0.95	0.8282828282828283	0.7374429223744292

AUROC and AUPRC curves and scores:





MLP (logistic) model (test)

	AUROC	AUPRC	Mean accuracy
MLP logistic model on test	0.81	0.61	0.74
On train	0.83	0.73	0.75

Compare with previous models on test data:

	Misclassification errors	AUROC	AUPRC	Mean accuracy
Logistic regression model	0.2125984251968504	0.83	0.71	0.79
SVM (linear model) model	0.21645021645021645	0.84	0.66	0.78
SVM (2nd degree poly kernel) model	0.2857142857142857	0.73	0.52	0.71
SVM (RBF kernel) model	0.22077922077922077	0.85	0.70	0.78
SVM (sigmoid kernel) model	0.2987012987012987	0.76	0.53	0.70
MLP logistic model	0.2554112554112554	0.81	0.61	0.74

Based on the above evaluation statistics, the MLP model with logistic activation function performs better than SVM models with 2nd degree polynomial kernel and sigmoid kernel, worse than the logistic regression model and SVM models with linear and RBF kernels. Here, no worry of overfitting because of small discrepancy between train and test errors.

Part b

Based on the model in part a:

- Different optimization procedures
  - lbfgs solver

  Confusion matrix on test:

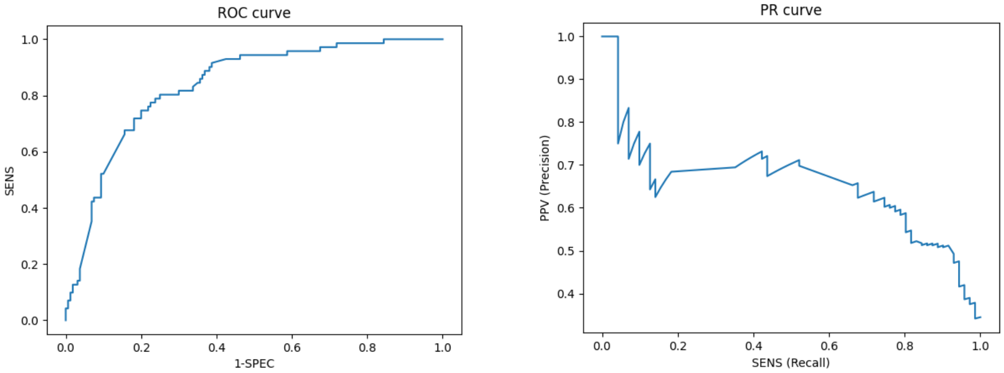
  ```
  [[131  29]
  [ 20  51]]
  ```

  On train:

  ```
  [[315  25]
   [ 63 134]]
  ```

	Misclassification error	SENS	SPEC	PPV	NPV
MLP (logistic, lbfgs) on test	0.21212121212121213	0.7183098591549296	0.81875	0.6375	0.8675496688741722
On train	0.16387337057728119	0.6802030456852792	0.9264705882352942	0.8427672955974843	0.8333333333333334

AUROC and AUPRC curves and scores on test:



MLP (logistic, lbfgs) model (test)

	AUROC	AUPRC	Mean accuracy
MLP logistic, lbfgs model on test	0.84	0.66	0.79
On train	0.90	0.82	0.84

Compared with the MLP with the default ‘adam’ solver, the model with ‘lbfgs’ has better performance with higher AUROC, AUPRC, mean accuracy, and smaller misclassification error. Its performance is similar to the SVM model with the linear kernel. Although the discrepancy between train and test errors is a little larger, it’s still small enough to indicate there is no overfitting problem.

- Change the number of iterations
    - 500 iterations

Confusion matrix on test:

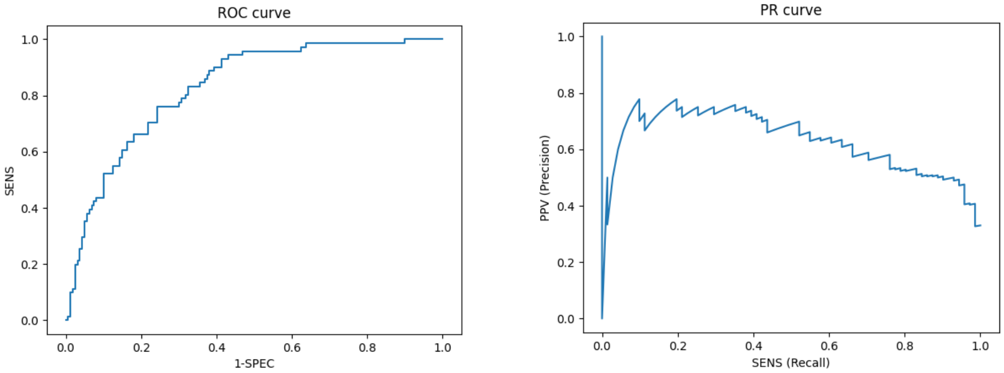
```
[[130  30]
 [ 24  47]]
```

On train:

```
[[298  42]
 [ 76 121]]
```

	Misclassification error	SENS	SPEC	PPV	NPV
MLP (logistic, 500 iter) on test	0.23376623376623376	0.6619718309859155	0.8125	0.6103896103896104	0.8441558441558441
On train	0.21973929236499068	0.6142131979695431	0.8764705882352941	0.7423312883435583	0.7967914438502673

AUROC and AUPRC curves and scores on test:



MLP (logistic, 500 iterations) model (test)

	AUROC	AUPRC	Mean accuracy
MLP logistic, 500 iter model on test	0.83	0.63	0.77
On train	0.84	0.74	0.78

Don’t need to worry about overfitting problems.

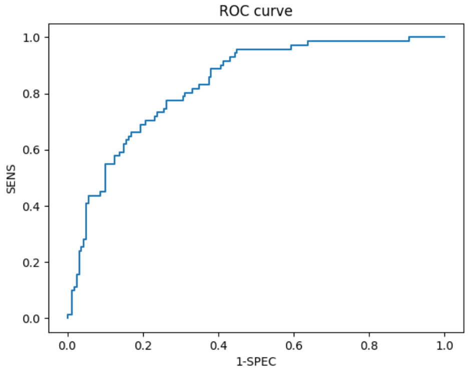
- 1000 iterations

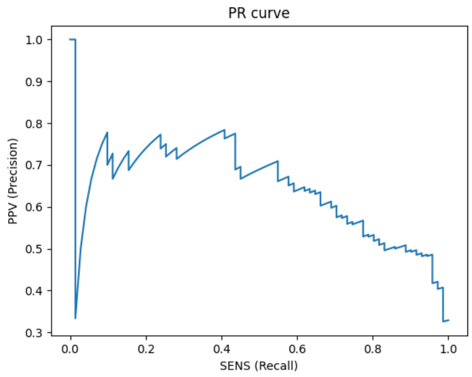
Confusion matrix on test:
[[127 33]
[22 49]]

On train:
[[293 47]
[70 127]]

	Misclassification error	SENS	SPEC	PPV	NPV
MLP (logistic, 1000) on test	0.23809523809523808	0.6901408450704225	0.79375	0.5975609756097561	0.8523489932885906
On train	0.21787709497206703	0.6446700507614214	0.861764705882353	0.7298850574712644	0.8071625344352618

AUROC and AUPRC curves and scores on test:





MLP (logistic, 1000 iterations) model (test)

	AUROC	AUPRC	Mean accuracy
MLP logistic, 1000 iter model on test	0.83	0.65	0.76
On train	0.84	0.75	0.78

Discrepancy is small, no overfitting problems.

- 100 iterations

Confusion matrix on test:

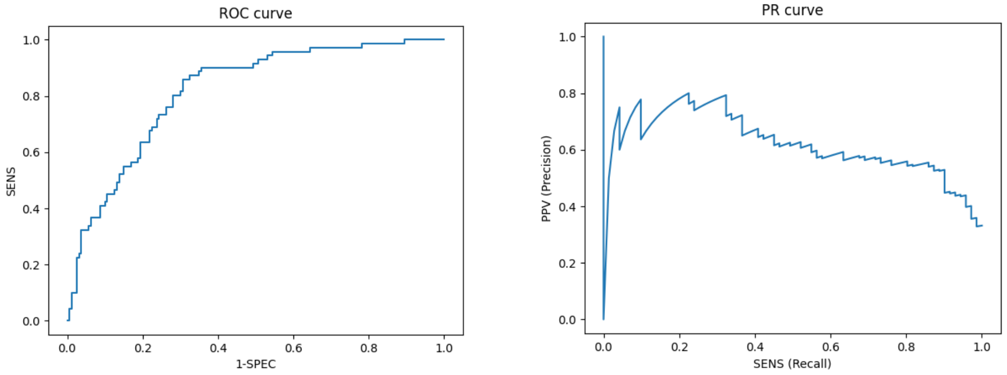
```
[[114  46]
 [ 14  57]]
```

On train:

```
[[258  82]
[ 60 137]]
```

	Misclassification error	SENS	SPEC	PPV	NPV
MLP (logistic, 100) on test	0.2597402597402597	0.8028169014084507	0.7125	0.5533980582524272	0.890625
On train	0.2644320297951583	0.6954314720812182	0.7588235294117647	0.6255707762557078	0.8113207547169812

AUROC and AUPRC curves and scores on test:



MLP (logistic, 100 iterations) model (test)

	AUROC	AUPRC	Mean accuracy
MLP logistic, 100 iter model on test	0.82	0.62	0.74
On train	0.82	0.72	0.74

No overfitting problems.

Here is a summary of the statistics for changing the number of iterations on test data:

	Misclassification errors	AUROC	AUPRC	Mean accuracy
MLP (logistic, 100)	0.2597402597402597	0.82	0.62	0.74
MLP logistic model (200 iter)	0.2554112554112554	0.81	0.61	0.74
MLP (logistic, 500 iter)	0.23376623376623376	0.83	0.63	0.77
MLP (logistic, 1000)	0.23809523809523808	0.83	0.65	0.76

Comparing the results after changing the number of iterations to different values, we can see that increasing the number of iterations improves the performance at first, but later the change in performance becomes more nuanced and the misclassification error even increases.

- Different units in 1 layer (the number of hidden layers is 1)
  - 3 units per layer

Confusion matrix on test:

```
[[136  24]
 [ 41  30]]
```

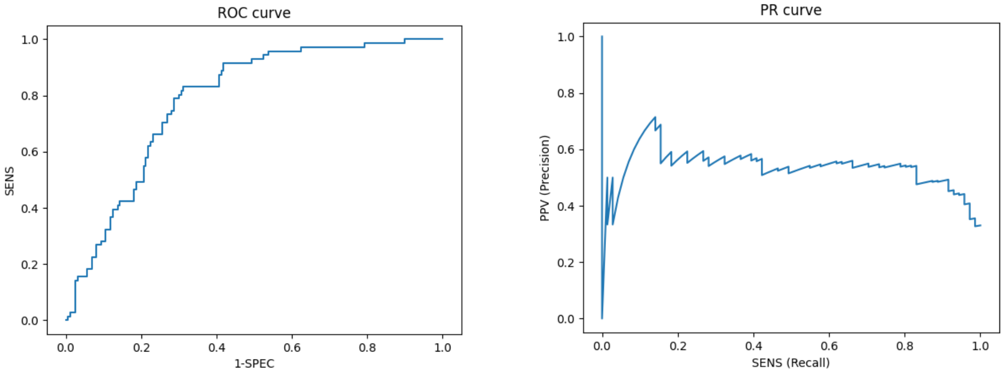
On train:

```
[[313  27]
 [118  79]]
```

	Misclassification error	SENS	SPEC	PPV	NPV

MLP (logistic, 3 units) on test	0.2813852813852814	0.4225352112676056	0.85	0.5555555555555556	0.768361581920904
On train	0.27001862197392923	0.4010152284263959	0.9205882352941176	0.7452830188679245	0.7262180974477959

AUROC and AUPRC curves and scores on test:



MLP (logistic, 3 units per layer) model (test)

	AUROC	AUPRC	Mean accuracy
MLP logistic, 3 units on test	0.78	0.54	0.72
On train	0.82	0.70	0.73

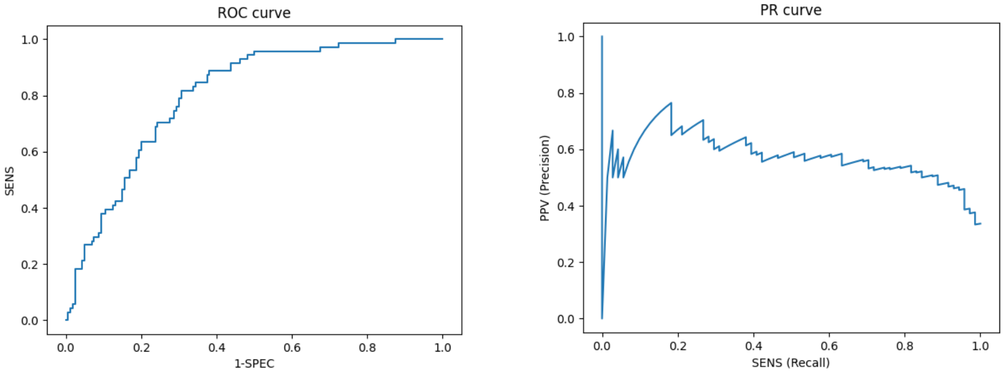
No overfitting problems because the test and train errors do not have a large difference.

- 5 units per layer

Confusion matrix on test:
[[136 24]
[41 30]]
On train:
[[313 27]
[109 88]]

	Misclassification error	SENS	SPEC	PPV	NPV
MLP (logistic, 5 units) on test	0.2813852813852814	0.4225352112676056	0.85	0.5555555555555556	0.768361581920904
On train	0.2532588454376164	0.4467005076142132	0.9205882352941176	0.7652173913043478	0.7417061611374408

AUROC and AUPRC curves and scores on test:



MLP (logistic, 5 units per layer) model (test)

	AUROC	AUPRC	Mean accuracy
MLP logistic, 5 units on test	0.80	0.58	0.73
On train	0.83	0.72	0.75

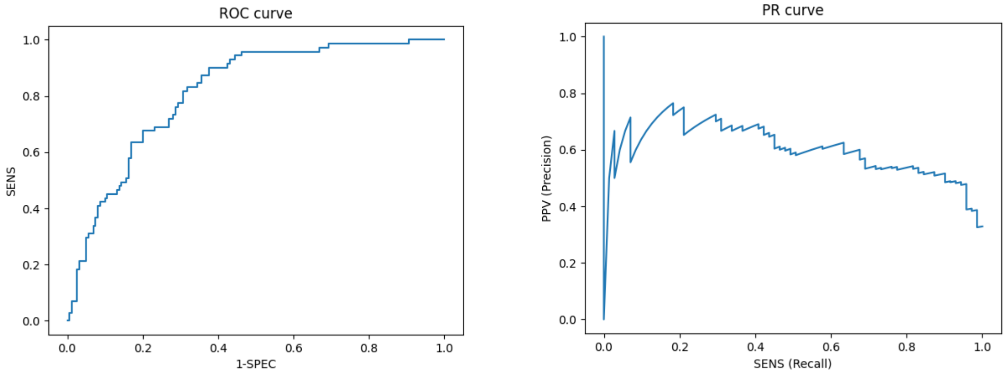
No need to worry about overfitting problems.

- 6 units

Confusion matrix on test:
[[130 30]
[26 45]]
On train:
[[299 41]
[85 112]]

	Misclassification error	SENS	SPEC	PPV	NPV
MLP (logistic, 6 units) on test	0.24242424242424243	0.6338028169014085	0.8125	0.6	0.8333333333333334
On train	0.2346368715083799	0.5685279187817259	0.8794117647058823	0.7320261437908496	0.7786458333333334

AUROC and AUPRC curves and scores on test:



MLP (logistic, 6 units per layer) model (test)

	AUROC	AUPRC	Mean accuracy
MLP logistic, 6 units on test	0.81	0.61	0.76
On train	0.84	0.74	0.77

There is no worry of overfitting problems.

Here is a summary of the statistics of changing the number of units per layer on test data:

	Misclassification errors	AUROC	AUPRC	Mean accuracy
MLP (logistic, 3 units)	0.2813852813852814	0.78	0.54	0.72
MLP logistic model (4 units)	0.2554112554112554	0.81	0.61	0.74
MLP (logistic, 5 units)	0.2683982683982684	0.80	0.58	0.73
MLP (logistic, 6 units)	0.24242424242424243	0.81	0.61	0.76

Although there are some exceptions (for example, 5 units per layer model performs worse than the 4 units one), overall, increasing the number of units per layer improves the performance in MLP models.

- Different number of hidden layers
  - 2 layers (4 units in first layer, 3 units in second)

Confusion matrix on test:

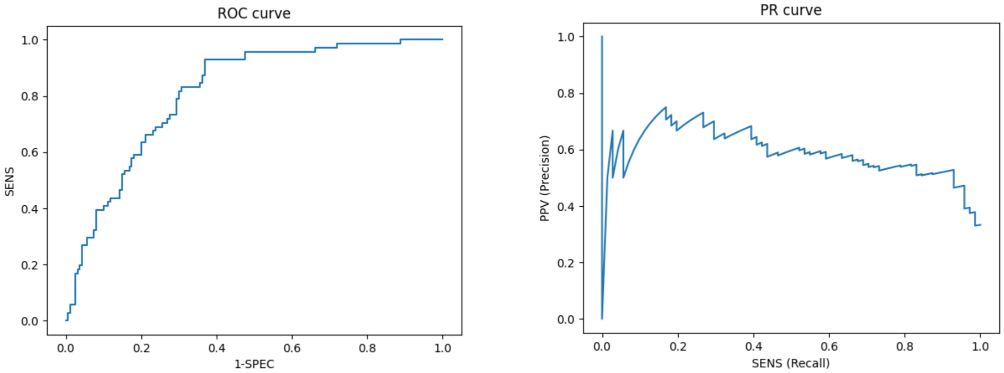
```
[[142  18]
 [ 42  29]]
```

On train:

```
[[321  19]
 [117  80]]
```

	Misclassification error	SENS	SPEC	PPV	NPV
MLP (logistic, 2 layers) on test	0.2597402597402597	0.4084507042253521	0.8875	0.6170212765957447	0.7717391304347826
On train	0.2532588454376164	0.40609137055837563	0.9441176470588235	0.8080808080808081	0.7328767123287672

AUROC and AUPRC curves and scores on test:



MLP (logistic, 2 hidden layers) model (test)

	AUROC	AUPRC	Mean accuracy
MLP logistic, 2 layers on test	0.81	0.60	0.74
On train	0.83	0.73	0.75

There is no worry for overfitting problems since the difference between train error and test error is very small.

Here, the performance of adding 1 more hidden layer is similar to the 1 hidden layer model.