Logistic Regression

Thomas Schwarz

Categorical Data

- Outcomes can be categorical
  - Often, outcome is binary:
    - President gets re-elected or not
    - Customer is satisfied or not
  - Often, explanatory variables are categorical as well
    - Person comes from an under-performing school
    - Order was made on a week-end
    - …

Prediction Models for Binary Outcomes

- Famous example:
  - Taken an image of a pet, predict whether this is a cat or a dog

Prediction Models for Binary Outcomes

- Bayes: ***generative classifier***
  - Predicts indirectly $P(c \mid d)$
  - $\hat{c} =$ arg max$_{c \in C} P(d \mid c) P(c)$

    Likelihood Prior

  - Evaluates product of likelihood and prior
    - Prior: Probability of a category c without looking at data
    - Likelihood: Probability of observing data if from a category c

Prediction Models for Binary Outcomes

- Regression is a ***discriminative classifier***
  - Tries to learn directly the classification from data
    - E.g.: All dog pictures have a collar
      - Collar present —> predict dog
      - Collar not present —> predict cat
  - Computes directly $P(c \mid d)$

Prediction Models for Binary Outcomes

- Regression:
  - Supervised learning: Have a training set with classification provided
  - Input is given as vectors of numerical features
    - $\mathbf{x}^{(i)} = (x_{1,i}, x_{2,i}, \ldots, x_{n,i})$
  - Classification function that calculates the predicted class $\hat{y}(\mathbf{x})$
  - An objective function for learning: Measures the goodness of fit between true outcome and predicted outcome
  - An algorithm to optimize the objective function

Prediction Models for Binary Outcomes

- Linear Regression:
  - Classification function of type
    - $\hat{y}\left((x_1, x_2, \ldots, x_n)\right) = a_1x_1 + a_2x_2 + \ldots a_nx_n + b$
  - Objective function (a.k.a cost function)
    - Sum of squared differences between predicted and observed outcomes
      - E.g. Test Set $T = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots \mathbf{x}^{(m)}\}$
      - Minimize cost function $\sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)})^2$

Prediction Models for Binary Outcomes

- Linear regression can predict a numerical value
  - It can be made to predict a binary value
    - If the predictor is higher than a cut-off value: predict yes
    - Else predict no
- But there are better ways to generate a binary classifier

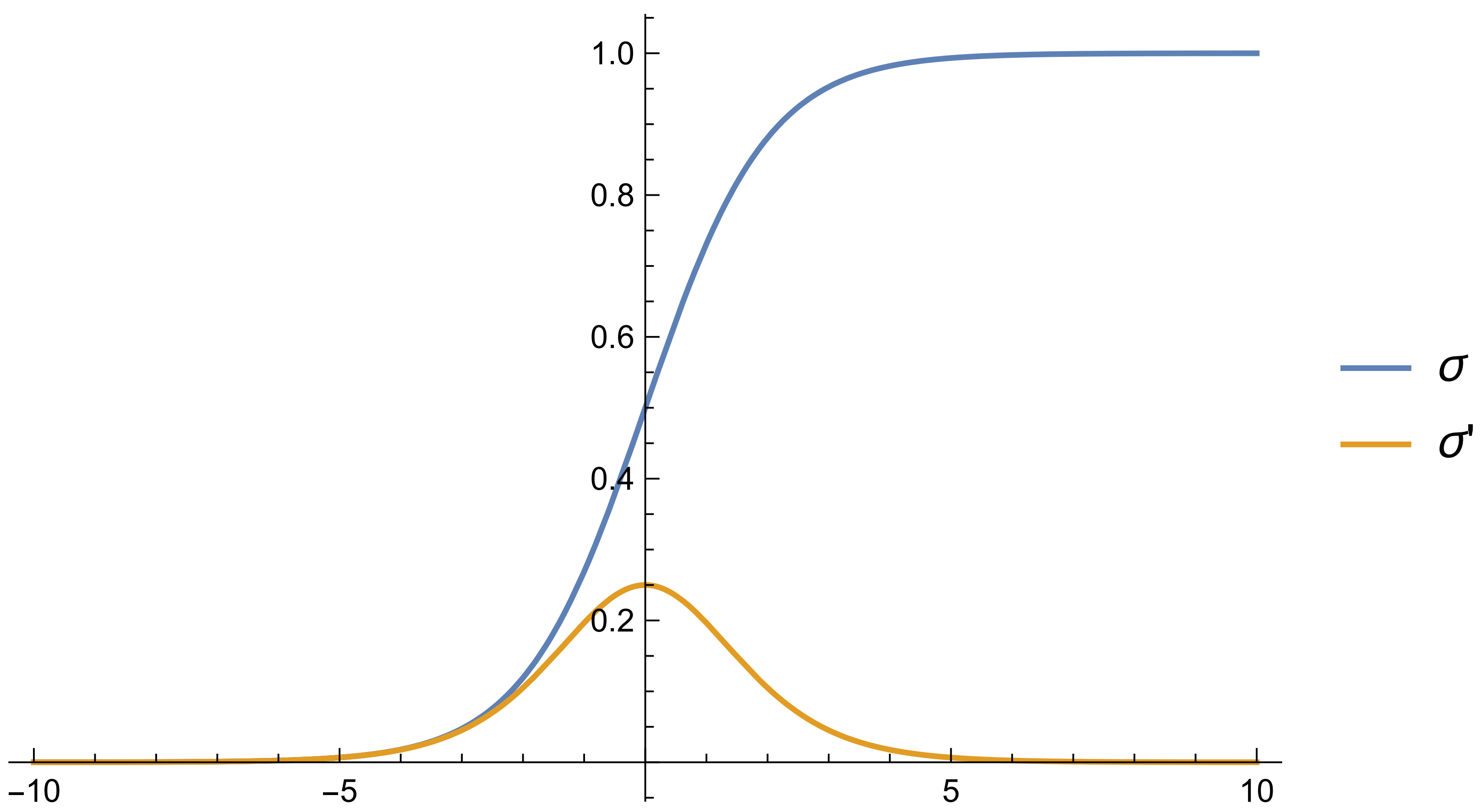
Prediction Models for Binary Outcomes

- Good binary classifier:
  - Since we want to predict the probability of a category based on the features:
    - Should look like a probability
  - Since we want to optimize:
    - Should be easy to differentiate
  - Best candidate classifier that has emerged:
    - Sigmoid classifier

Logistic Regression

- Use logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$



Logistic Regression

- Combine with linear regression to obtain logistic regression approach:
  - Learn best weights in
    - $\hat{y}\left((x_1, x_2, \ldots, x_n)\right) = \sigma(b + w_1 x_1 + w_2 x_2 + \ldots w_n x_n)$
  - We know interpret this as a probability for the positive outcome '+'
  - Set a ***decision boundary*** at 0.5
    - This is no restriction since we can adjust b and the weights

Logistic Regression

- We need to measure how far a prediction is from the true value
  - Our predictions $\hat{y}$ and the true value y can only be 0 or 1
    - If $y = 1$: Want to support $\hat{y} = 1$ and penalize $\hat{y} = 0$.
    - If $y = 0$: Want to support $\hat{y} = 0$ and penalize $\hat{y} = 1$.
  - One successful approach:
    - Loss$(\hat{y}, y) = \hat{y}^y (1 - \hat{y})^{(1-y)}$

Logistic Regression

- Easier: Take the negative logarithm of the loss function
  - Cross Entropy Loss

$$L_{CE} = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$

Logistic Regression

- This approach is successful, because we can use Gradient Descent
  - Training set of size m
  - Minimize $\sum_{i=1}^{m} \mathrm{L}_{\mathrm{CE}}(y^{(i)}, \hat{y}^{(i)})$
  - Turns out to be a convex function, so minimization is simple! (As far as those things go)
  - Recall:

$$\hat{y}\left((x_1, x_2, \ldots, x_n)\right) = \sigma(b + w_1 x_1 + w_2 x_2 + \ldots w_n x_n)$$

  - We minimize with respect to the weights and b

Logistic Regression

- Calculus:

$$\frac{\delta \mathsf{L}_{\mathsf{CE}}(w, b)}{\delta w_j} = \big(\sigma(w_1 x_1 + \ldots w_n x_n + b) - y\big)\, x_j$$

$$= (\hat{y} - y) x_j$$

- Difference between true y and estimated outcome $\hat{y}$, multiplied by input coordinate

Logistic Regression

- Stochastic Gradient Descent
  - Until gradient is almost zero:
    - For each training point $x^{(i)}, y^{(i)}$:
      - Compute prediction $\hat{y}^{(i)}$
      - Compute loss
      - Compute gradient
      - Nudge weights in the opposite direction using a learning weight η
        - $(w_1, \ldots, w_n) \leftarrow (w_1, \ldots, w_n) - \eta \nabla \mathrm{L}_{\mathrm{CE}}$
      - Adjust η

Logistic Regression

- Stochastic gradient descent uses a single data point
  - Better results with random batches of points at the same time

Lasso and Ridge Regression

- If the feature vector is long, danger of overfitting is high
  - We learn the details of the training set
    - Want to limit the number of features with positive weight
  - Dealt with by adding a regularization term to the cost function
    - Regularization term depends on the weights
      - Penalizes large weights

Lasso and Ridge Regression

- L2 regularization:
  - Use a quadratic function of the weights
    - Such as the euclidean norm of the weights
  - Called ***Ridge Regression***
    - Easier to optimize

Lasso and Ridge Regression

- L1 regularization
  - Regularization term is the sum of the absolute values of weights
  - Not differentiable, so optimization is more difficult
  - BUT: effective at lowering the number of non-zero weights
- Feature selection:
  - Restrict the number of features in a model
  - Usually gives better predictions

Examples

- Example: quality.csv
  - Try to predict whether patient labeled care they received as poor or good

quality

MemberID	InpatientDays	ERVisits	OfficeVisits	Narcotics	DaysSinceLastERVisit	Pain	TotalVisits	ProviderCount	MedicalClaims	ClaimLines	StartedOnCombination	AcuteDrugGapSmall	PoorCare
1	0	0	18	1	731	10	18	21	93	222	FALSE	0	0
2	1	1	6	1	411	0	8	27	19	115	FALSE	1	0
3	0	0	5	3	731	10	5	16	27	148	FALSE	5	0
4	0	1	19	0	158	34	20	14	59	242	FALSE	0	0
5	8	2	19	3	449	10	29	24	51	204	FALSE	0	0
6	2	0	9	2	731	6	11	40	53	156	FALSE	4	1
7	16	1	8	1	173.9583333	4	25	19	40	261	FALSE	0	0
8	2	0	8	0	731	5	10	11	28	87	FALSE	0	0
9	2	1	4	3	45	5	7	28	20	98	FALSE	0	1
10	4	2	0	2	104	2	6	21	17	66	FALSE	0	0
11	6	5	20	2	156	9	31	19	43	126	FALSE	2	0
12	0	0	7	4	731	0	7	8	23	41	FALSE	2	0
13	0	1	3	1	389	23	4	13	18	70	FALSE	0	0
14	1	1	20	3	594.9583333	16	22	18	48	133	FALSE	0	0
15	6	2	31	3	640.9583333	70	39	28	101	233	FALSE	0	0
16	0	0	8	0	731	0	8	5	19	48	FALSE	0	0
17	2	0	9	0	731	29	11	22	39	120	FALSE	0	0
18	3	0	20	1	731	13	23	17	34	73	FALSE	3	1

Examples

- First column is an arbitrary patient ID
  - we make this the index
- One column is a Boolean, when imported into Python
  - so we change it to a numeric value

```
df = pd.read_csv('quality.csv', sep=',', index_col=0)
df.replace({False:0, True:1}, inplace=True)
```

Examples

- Farmington Heart Data Project:
  - https://framinghamheartstudy.org
  - Monitoring health data since 1948
  - 2002 enrolled grandchildren of first study

Examples

framingham

male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
1	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77	0
0	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76	0
1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0
0	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103	1
0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0
0	43	2	0	0	0	0	1	0	228	180	110	30.3	77	99	0
0	63	1	0	0	0	0	0	0	205	138	71	33.11	60	85	1
0	45	2	1	20	0	0	0	0	313	100	71	21.68	79	78	0
1	52	1	0	0	0	0	1	0	260	141.5	89	26.36	76	79	0
1	43	1	1	30	0	0	1	0	225	162	107	23.61	93	88	0
0	50	1	0	0	0	0	0	0	254	133	76	22.91	75	76	0

Examples

- Contains a few NaN data
  - We just drop them

```
df = pd.read_csv('framingham.csv', sep=',')
df.dropna(inplace=True)
```

Logistic Regression in Stats-Models

- Import statsmodels.api

```
import statsmodels.api as sm
```

- Interactively select the columns that gives us high p-values

```
cols = [ 'Pain', 'TotalVisits',
         'ProviderCount',
         'MedicalClaims', 'ClaimLines',
         'StartedOnCombination',
         'AcuteDrugGapSmall',]
```

Logistic Regression in Stats-Models

- Create a logit model
  - Can do as we did for linear regression with a string
  - Can do using a dataframe syntax

```
logit_model=sm.Logit(df.PoorCare,df[cols])
result=logit_model.fit()
```

  - Print the summary pages

```
print(result.summary2())
```

Logistic Regression in Stats-Models

- Print the results
  - `print(result.pred_table())`
- This gives the "confusion matrix"
  - Coefficient [i,j] gives:
    - predicted i values
    - actual j values

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Logistic Regression in Stats-Models

- Quality prediction:
  - ```
    [[91.  7.]
     [18. 15.]]
    ```
- 7 False negative and 18 false positives
    ```

Logistic Regression in Stats-Models

- Heart Event Prediction:
  - ```
    [[3075.   26.]
     [ 523.   34.]]
    ```

- 26 false negatives
- 523 false positives
    ```

Logistic Regression in Stats-Models

- Can try to improve using Lasso

```
result=logit_model.fit_regularized()
```

Logistic Regression in Stats-Models

- Can try to improve selecting only columns with high P-values

```
Optimization terminated successfully.
         Current function value: 0.423769
         Iterations 6
                        Results: Logit
================================================================
Model:              Logit            Pseudo R-squared: 0.007
Dependent Variable: TenYearCHD       AIC:              3114.2927
Date:               2020-07-12 18:18 BIC:              3157.7254
No. Observations:   3658             Log-Likelihood:   -1550.1
Df Model:           6                LL-Null:          -1560.6
Df Residuals:       3651             LLR p-value:      0.0019166
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
----------------------------------------------------------------
                  Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
----------------------------------------------------------------
currentSmoker     0.0390   0.0908  0.4291 0.6679 -0.1391  0.2170
BPMeds            0.5145   0.2200  2.3388 0.0193  0.0833  0.9457
prevalentStroke   0.7716   0.4708  1.6390 0.1012 -0.1511  1.6944
prevalentHyp      0.8892   0.0983  9.0439 0.0000  0.6965  1.0818
diabetes          1.4746   0.2696  5.4688 0.0000  0.9461  2.0030
totChol          -0.0067   0.0007 -9.7668 0.0000 -0.0081 -0.0054
glucose          -0.0061   0.0019 -3.2113 0.0013 -0.0098 -0.0024
================================================================
```

Logistic Regression in Stats-Models

- Select the columns
  - ```
    cols = ['currentSmoker', 'BPMeds',
            'prevalentStroke', 'prevalentHyp',
            'diabetes', 'totChol','glucose']
    ```
- Get a better (?) confusion matrix:
  - ```
    [[3086.   15.]
     [ 549.    8.]]
    ```
  - False negatives has gone down
  - False positives has gone up

Logistic Regression in Scikit-learn

- Import from sklearn

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
```

Logistic Regression in Scikit-learn

- Create a logistic regression object and fit it on the data

```
logreg = LogisticRegression()
logreg.fit(X=df[cols], y = df.TenYearCHD)
y_pred = logreg.predict(df[cols])
confusion_matrix = confusion_matrix(df.TenYearCHD,
y_pred)
print(confusion_matrix)
```

Logistic Regression in Scikit-learn

- Scikit-learn uses a different algorithm
  - Confusion matrix on the whole set is
    -

```
[[3087   14]
 [ 535   22]]
```

Logistic Regression in Scikit-learn

- Can also divide the set in training and test set

```
X_train, X_test, y_train, y_test =
   train_test_split(df[cols],
                    df.TenYearCHD,
                    test_size=0.3,
                    random_state=0)
logreg.fit(X_train, y_train)
y_pred = logreg.predict(X_test)
confusion_matrix = confusion_matrix(y_test, y_pred)
print(confusion_matrix)
```

Logistic Regression in Scikit-learn

- Confusion matrix
  - ```
    [[915   1]
     [176   6]]
    ```
    ```

Measuring Success

	precision	recall	f1-score	support
0	0.84	1.00	0.91	916
1	0.86	0.03	0.06	182
accuracy			0.84	1098
macro avg	0.85	0.52	0.49	1098
weighted avg	0.84	0.84	0.77	1098

Measuring Success

- How can we measure accuracy?
  - accuracy = (fp+fn)/(tp+tn+fp+fn)
    - Unfortunately, because of skewed data sets, often very high
  - precision = tp/(tp+fp)
  - recall = tp/(tp+fn)
  - F measure = harmonic mean of precision and recall

Probit Regression

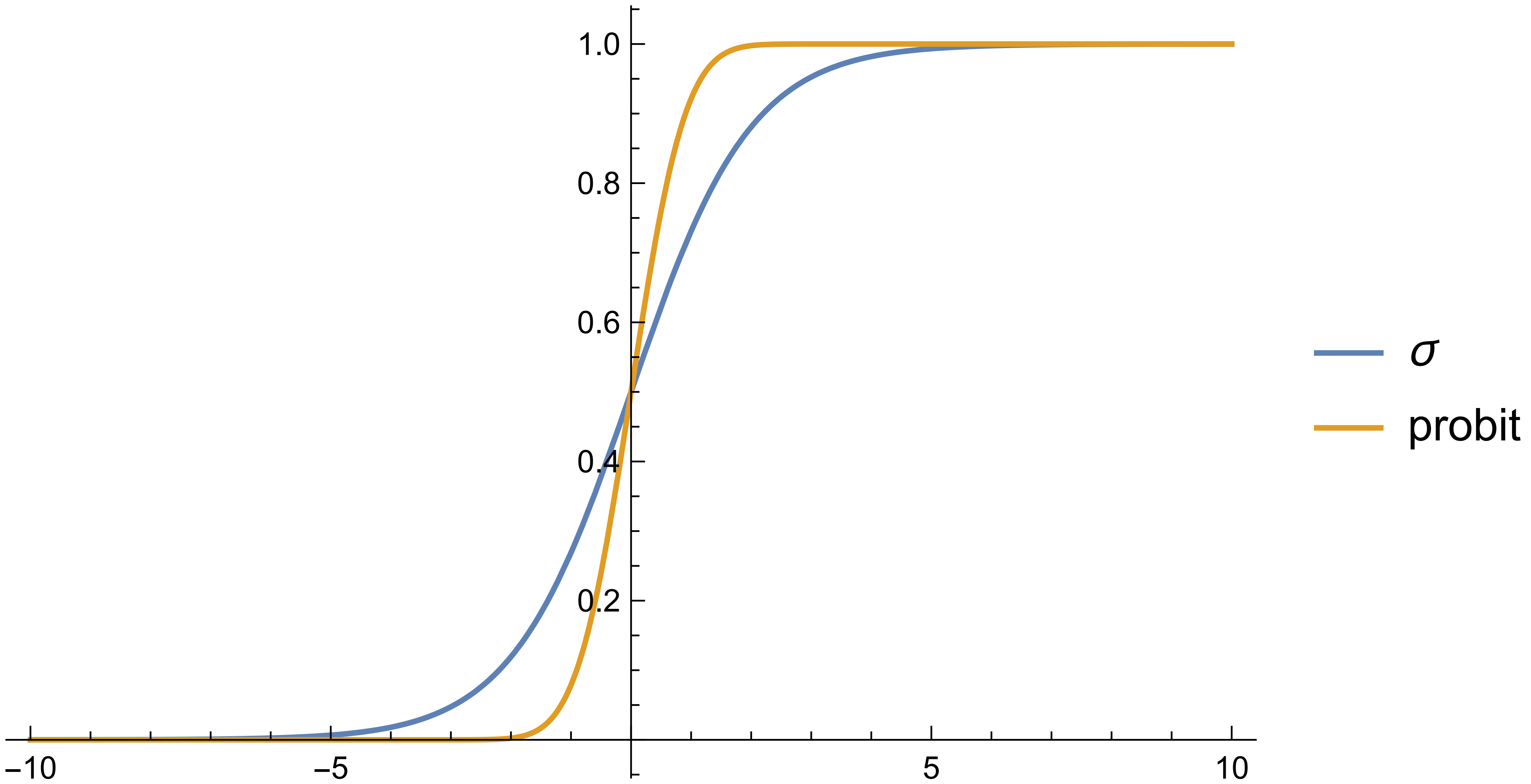
- Instead of using the logistic function σ, can also use the cumulative distribution function of the normal distribution

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp(-t^2)dt$$

- Predictor is then

$$\frac{1}{2}\left(1 + \text{erf}(b + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n)\right)$$

Probit Regression



Probit Regression

- Calculations with probit are more involved
  - Statsmodels implements it
    - `from statsmodels.discrete.discrete_model import Probit`
  - Fit the probit model

```
probit_model=Probit(df.TenYearCHD,df[cols])
result=probit_model.fit()

print(result.summary())
print(result.pred_table())
for i in range(20):
   print(df.TenYearCHD.iloc[i],
result.predict(df[cols]).iloc[i])
```

Probit Regression

- Confusion matrix is now

```
[[3085.   16.]
 [ 547.   10.]]
```

- More false positives but less false negatives

Multinomial Logistic Regression

- Want to predict one of several categories based on feature vector
- Use the softmax function

$$\text{softmax}\left(z_1, z_2, \ldots, z_m\right) = \left(\frac{e^{z_1}}{\sum_{i=1}^{m} e^{z_i}}, \frac{e^{z_2}}{\sum_{i=1}^{m} e^{z_i}}, \ldots, \frac{e^{z_m}}{\sum_{i=1}^{m} e^{z_i}}\right)$$

Multinomial Logistic Regression

- Learning is still possible, but more complicated

Multinomial Logistic Regression

```
model1 = LogisticRegression(random_state=0,
                            multi_class='multinomial',
                            penalty='none',
                            solver='newton-cg').fit(X_train, y_train)
preds = model1.predict(X_test)
```