CS 2750 Machine Learning
Lecture 11

Evaluation of classifiers

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Classification model learning

Learning:

- Many different ways and objective criteria used to learn the classification models. Examples:
  - Mean squared errors to learn the discriminant functions
  - Negative log likelihood (logistic regression)

Evaluation:

- One possibility: Use the same error criteria as used during the learning (apply to train & test data). Problems:
  - May work for discriminative models
  - Harder to interpret for humans.
- **Question:** how to more naturally evaluate the classifier performance?

Evaluation of classification models

For any data set we use to test the classification model on we can build a **confusion matrix:**

– Counts of examples with:

– class label ω_j that are classified with a label α_i

target

		$\omega = 1$	$\omega = 0$
predict	$\alpha = 1$	140	17
	$\alpha = 0$	20	54

Evaluation of classification models

Confusion matrix entries are often normalized with respect to the number of examples N to get proportions of the different agreements and disagreements among predicted and target values

target

		$\omega = 1$	$\omega = 0$
predict	$\alpha = 1$	$140/231$	$17/231$
	$\alpha = 0$	$20/231$	$54/231$

Basic evaluation statistics

Basic statistics calculated from the confusion matrix:

target

		$\omega = 1$	$\omega = 0$
predict	$\alpha = 1$	140	17
	$\alpha = 0$	20	54

Classification Accuracy = 194/231

Basic evaluation statistics

Basic statistics calculated from the confusion matrix:

target

		$\omega = 1$	$\omega = 0$
predict	$\alpha = 1$	140	17
	$\alpha = 0$	20	54

Classification Accuracy = 194/231

Misclassificion Error = 37/231 = 1 - **Accuracy**

Evaluation for binary classification

Entries in the confusion matrix for binary classification have names:

target

		$\omega = 1$	$\omega = 0$
predict	$\alpha = 1$	TP	FP
	$\alpha = 0$	FN	TN

TP: True positive (hit)
FP: False positive (false alarm)
TN: True negative (correct rejection)
FN: False negative (a miss)

Additional statistics

- **Sensitivity (recall)**

$$SENS = \frac{TP}{TP + FN}$$

- **Specificity**

$$SPEC = \frac{TN}{TN + FP}$$

- **Positive predictive value (precision)**

$$PPT = \frac{TP}{TP + FP}$$

- **Negative predictive value**

$$NPV = \frac{TN}{TN + FN}$$

Binary classification: additional statistics

- **Confusion matrix**

		target		
		1	0	
predict	1	140	10	$PPV = 140/150$
	0	20	180	$NPV = 180/200$
		$SENS = 140/160$	$SPEC = 180/190$	

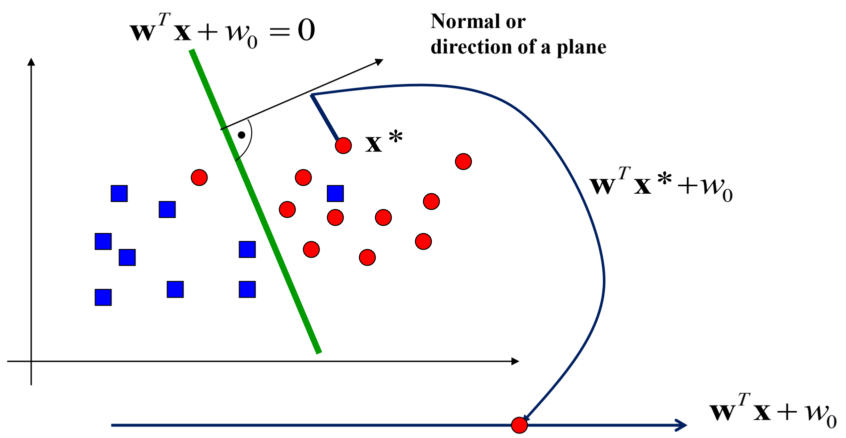
Row and column quantities:

- Sensitivity (SENS)
- Specificity (SPEC)
- Positive predictive value (PPV)
- Negative predictive value (NPV)

Binary classification models

Often project data points to one dimensional space:

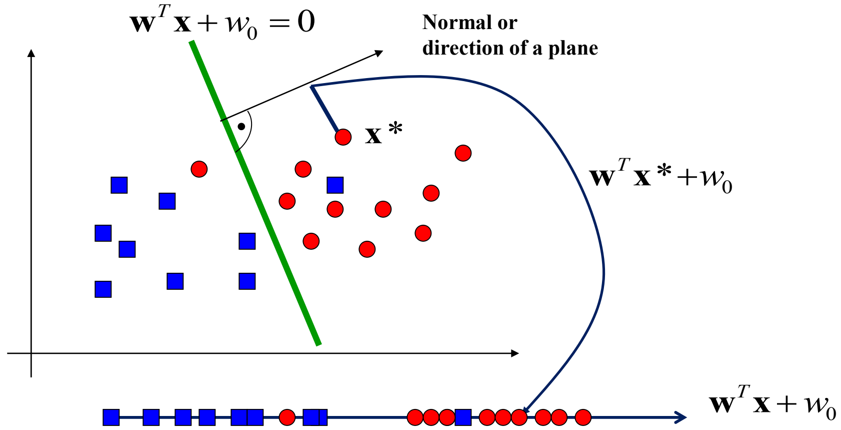
Defined for example by: $\mathbf{w}^T\mathbf{x}+w_0$ or $p(y=1|\mathbf{x},\mathbf{w})$



Binary classification models

Often project data points to one dimensional space:

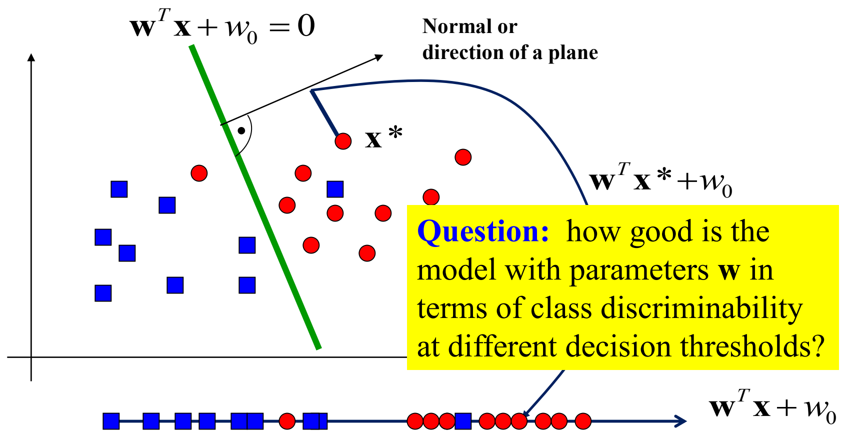
Defined for example by: $\mathbf{w}^T\mathbf{x}+w_0$ or p(y=1|**x**,**w**)



Binary classification models

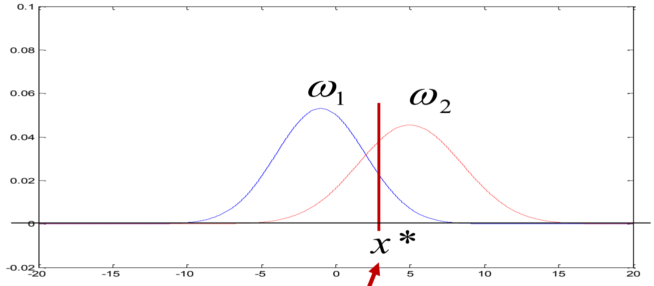
Often project data points to one dimensional space:

Defined for example by: $\mathbf{w}^T\mathbf{x}+w_0$ or p(y=1|**x**,**w**)



Question: how good is the model with parameters **w** in terms of class discriminability at different decision thresholds?

Receiver Operating Characteristic (ROC)



- **Probabilities:**
  - *SENS* $p(x > x^* \mid \mathbf{x} \in \omega_2)$
  - *SPEC* $p(x < x^* \mid \mathbf{x} \in \omega_1)$

threshold

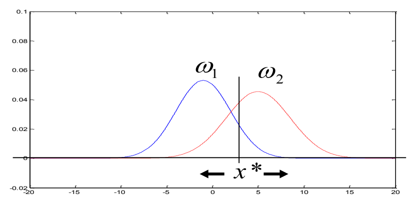
Receiver Operating Characteristic (ROC)

- **ROC curve plots :**

SN= $p(x > x^* \mid \mathbf{x} \in \omega_2)$

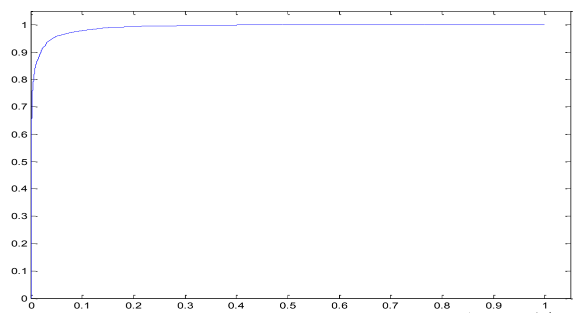
1-SP= $p(x > x^* \mid \mathbf{x} \in \omega_1)$

for different x*



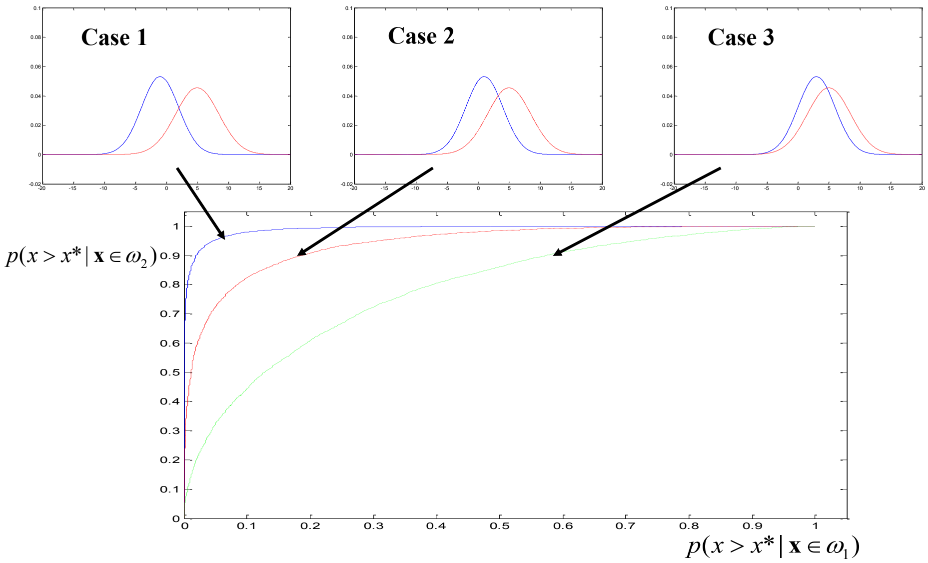
SENS

$p(x > x^* \mid \mathbf{x} \in \omega_2)$



1-SPEC $p(x > x^* \mid \mathbf{x} \in \omega_1)$

ROC curve



Receiver operating characteristic

- **ROC**
  - shows the discriminability between the two classes under different thresholds representing different decision biases
- **Decision bias**
  - can be changed using the different loss function
- **Quality of a classification model:**
  - Area under the ROC
  - Best value 1, worst (no discriminability): 0.5