Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 15: Support Vector Machines

Mar 14, 2016

classification, so far

Recall the perceptron

$$\hat{y}_i = \begin{cases} 1 & \text{if } \sum_i^F x_i\beta_i \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

Algorithm 4 Perceptron stochastic gradient descent

```
Data: training data x ∈ ℝ^F, y ∈ {−1, 1}
β = 0^F
η = 1                                        ▷ step size
while not converged do
    for i = 1 to N do
        β_{t+1} = β_t + η y_i x_i
    end for
end while
```

Recall the perceptron

- At the end of training, the coefficients β are a linear combination of the inputs x

$$\hat{\beta} = \sum_{i=1}^{N} \alpha_i y_i x_i$$

- α_i = the number of times data point i was misclassified

Recall the perceptron

$$\hat{y}_i = \begin{cases} 1 & \text{if } \beta^\top x_i \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

$$\hat{y}_i = \begin{cases} 1 & \text{if } \left(\sum_{j=1}^N \alpha_j y_j x_j\right)^\top x_i \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

$$\hat{y}_i = \begin{cases} 1 & \text{if } \sum_{j=1}^N \alpha_j y_j \left(x_j^\top x_i\right) \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

Recall the perceptron

$$\hat{y}_i = \begin{cases} 1 & \text{if } \sum_{j=1}^{N} \alpha_j y_j \left(x_j^\top x_i \right) \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

We can replace this inner product with a kernel

Kernels

$$\kappa(x,\ x') \in \mathbb{R}$$

- Often symmetric — $K(x', x) = K(x, x')$
- And non-negative — $K(x, x') \geq 0$ (but need not be)
- Often thought of as a measure of “similarity”

Kernels

dot product = linear kernel

$$\kappa(x,\ x') = x^\top x' = \sum_{i=1}^{F} x_i\, x_i'$$

cosine similarity kernel

$$\kappa(x,\ x') = \frac{\sum_{i=1}^{F} x_i\, x_i'}{\sqrt{\sum_{i=1}^{F} x_i^2}\sqrt{\sum_{i=1}^{F} x_i'^2}}$$

Kernels

Gaussian kernel/RBF kernel

$$\kappa(x,\ x') = \exp\left(-\frac{1}{2} \sum_{i=1}^{F} \frac{1}{\sigma_i^2} (x_i - x_i')^2 \right)$$

Higher dimensions

$$K(x, x') = (x^\top x')^2$$

$$= \left(\sum_{i=1}^{F} (x_i x_i') \right)^2$$

$$= \sum_{i=1}^{F} (x_i x_i') \sum_{i=j}^{F} (x_j x_j')$$

Higher dimensions

$$= \sum_{i=1}^{F} (x_i x_i') \sum_{i=j}^{F} (x_j x_j')$$

$$= \sum_{i=1}^{F} \sum_{j=1}^{F} x_i x_j x_i' x_j'$$

$$= \sum_{i,j=1}^{F} (x_i x_j)(x_i' x_j')$$

Higher dimensions

$$= \sum_{i,j=1}^{F} (x_i x_j)(x'_i x'_j)$$

$$= \phi(x)^\top \phi(x')$$

Non-linear kernels imply a higher-dimensional feature representation for x

$$\Phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}$$

"Implicit" feature space

$$(x^\top x')^2 = \phi(x)^\top \phi(x')$$

x	x'
x_1	x'_1
x_2	x'_2
x_3	x'_3

original feature space

$\phi(x)$	$\phi(x')$
x_1x_1	$x'_1x'_1$
x_1x_2	$x'_1x'_2$
x_1x_3	$x'_1x'_3$
x_2x_1	$x'_2x'_1$
x_2x_2	$x'_2x'_2$
x_2x_3	$x'_2x'_3$
x_3x_1	$x'_3x'_1$
x_3x_2	$x'_3x'_2$
x_3x_3	$x'_3x'_3$

implied feature space

x

good	1
not	1
movie	0

original feature space

$\phi(x)$

good good	1
good not	1
good movie	0
not good	1
not not	1
not movie	0
movie good	0
movie not	0
movie movie	0

implied feature space

Kernels

A

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29

B

so much depends
upon

a red wheel
barrow

glazed with rain
water

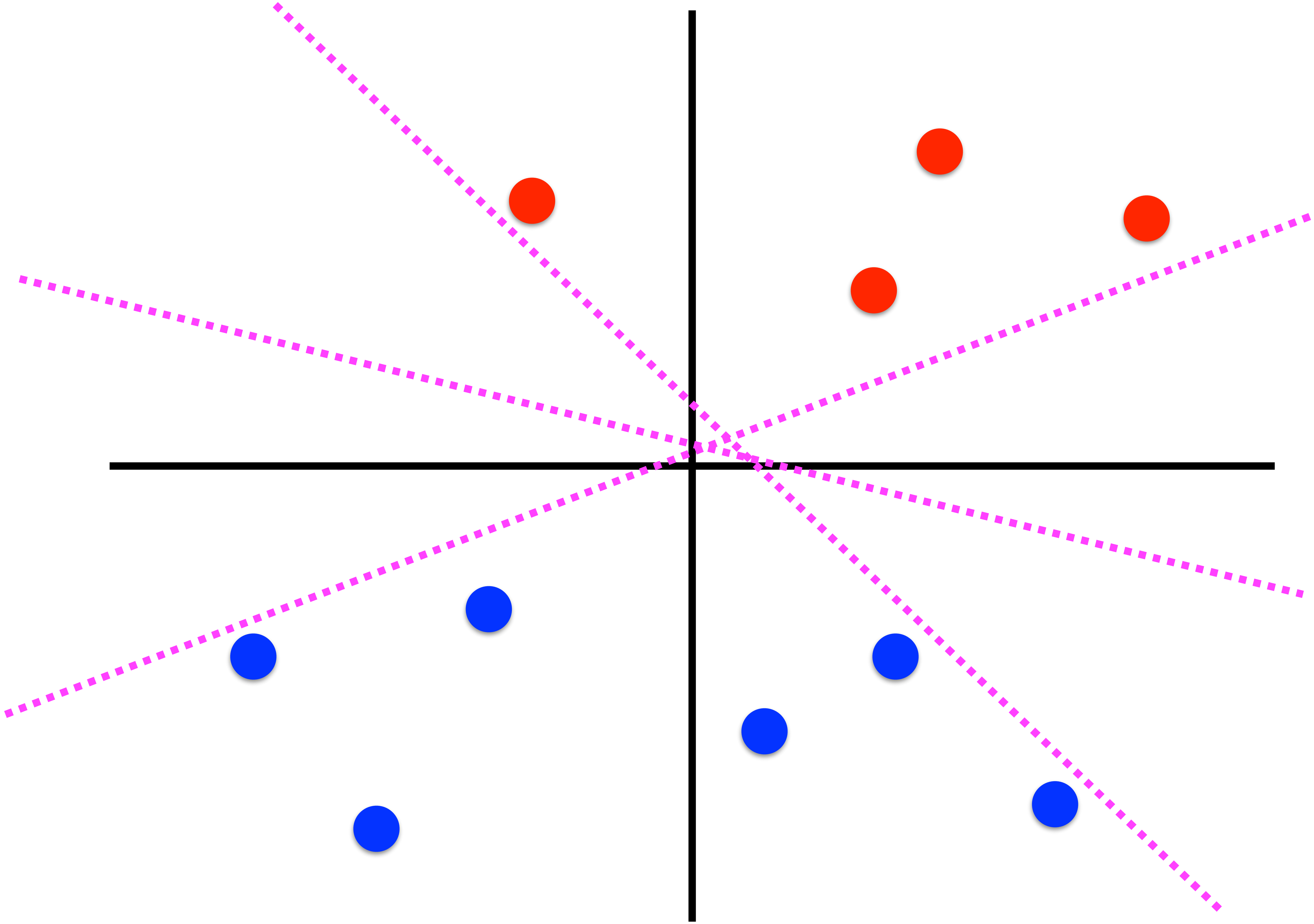
beside the white
chickens.

Code

Support vector machines

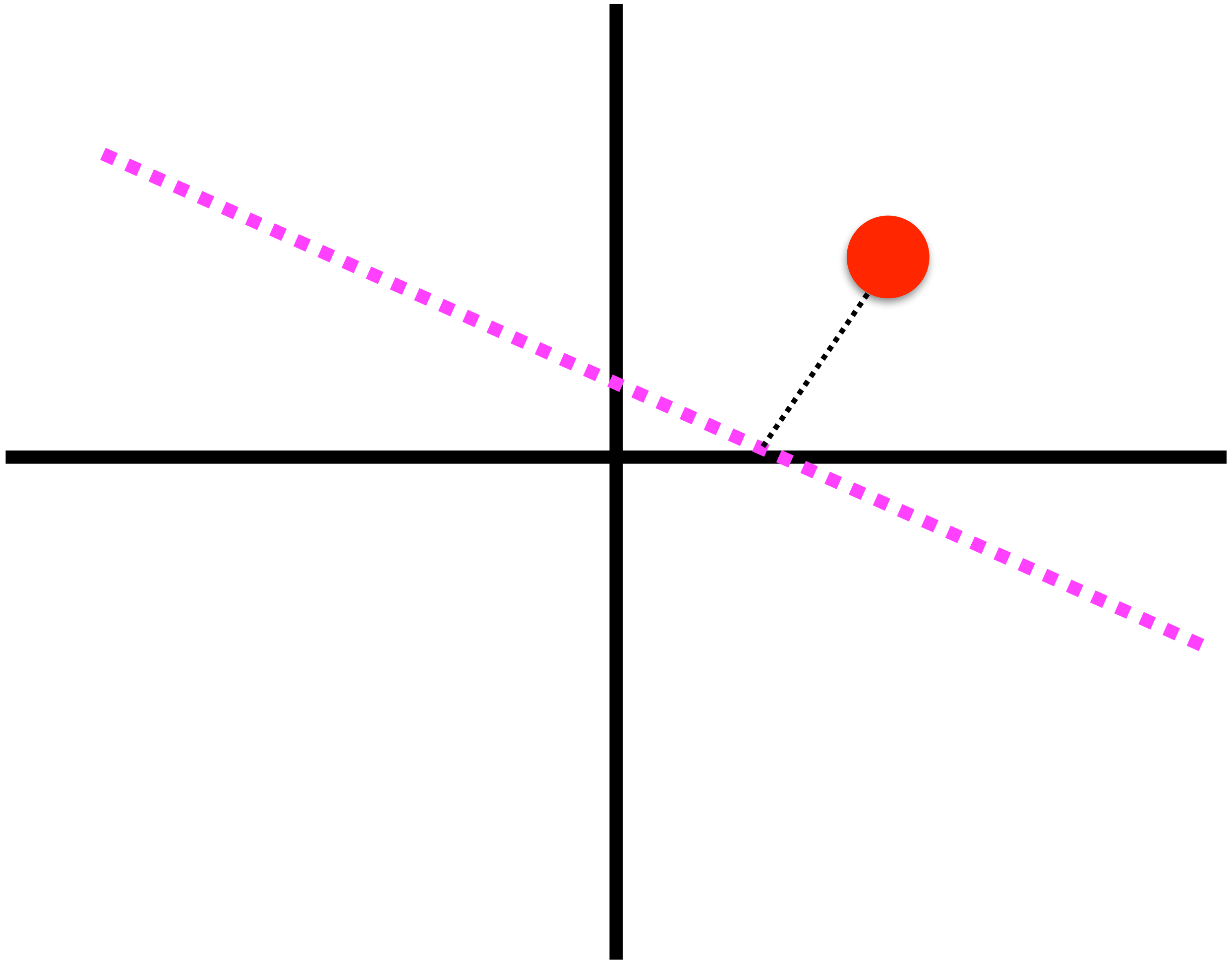
Two principles:

1. Kernel trick
2. Margin maximization



Margin

- Distance from the closest point to the decision boundary



Support vector machines

- For all of the training examples, we want to:
  - Maximize the margin
  - Subject to all of the training examples being on the correct side.

Loss functions

log loss
(logistic regression)

$$-\sum_{i=1}^{N} \log P(y \mid x, \beta) - \sum_{j=1}^{F} \beta_j^2$$

hinge loss
(SVM)

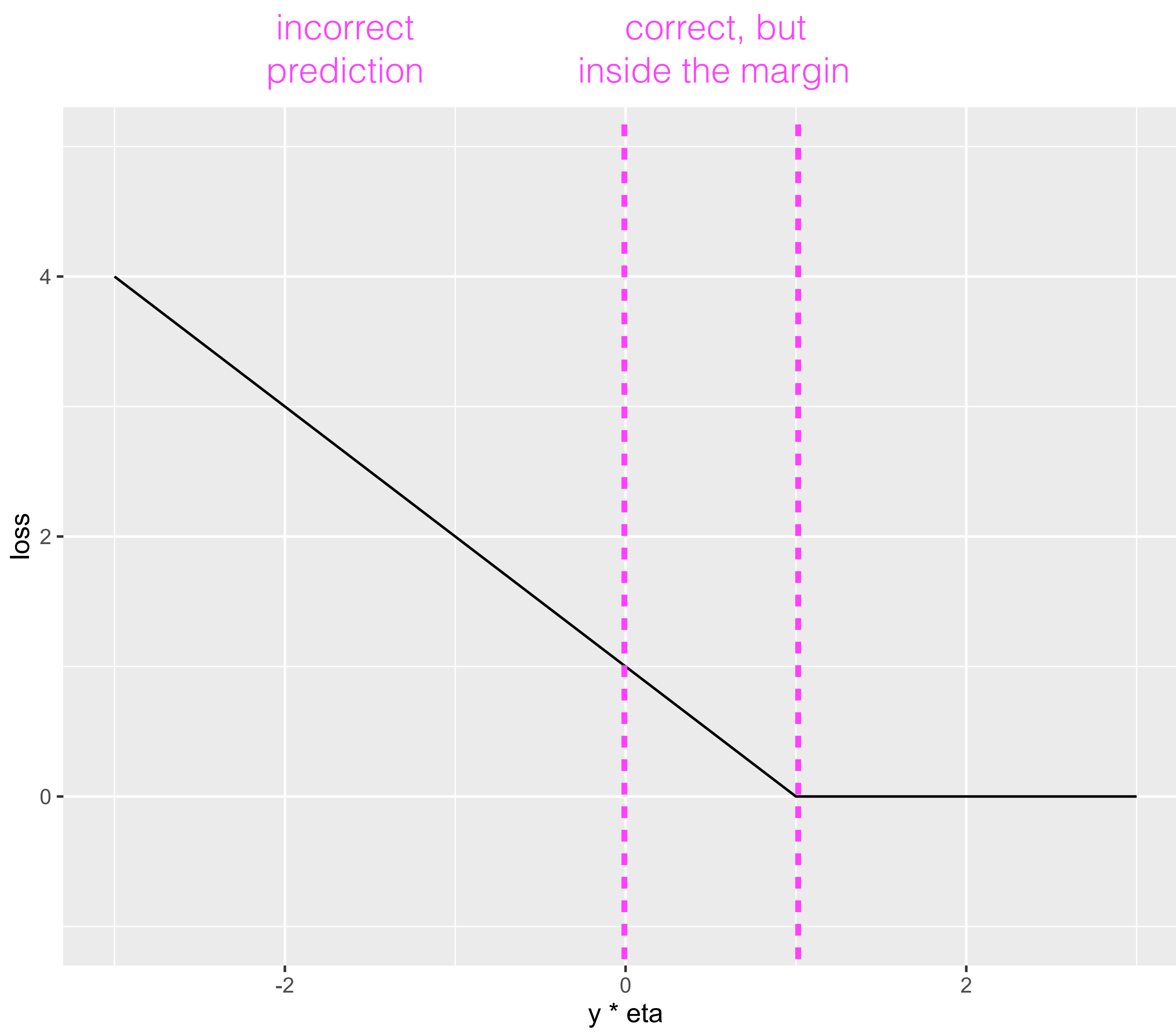
$$\sum_{i=1}^{N} \max(0, 1 - y\eta) - \sum_{j=1}^{F} \beta_j^2$$

No loss is suffered if the prediction is outside the margin on the correct side

Hinge loss

$$\max(0, 1 - y\eta)$$

η = score
$y = \{1, -1\}$



Support vector machines

"slack variable" $\xi_i = max(0, 1 - y_i \eta_i)$

$$\arg\min_{\beta} \; \overbrace{C\frac{1}{n}\sum_{i=1}^{N}\xi_i}^{\text{loss}} + \overbrace{\sum_{j=1}^{F}\beta_j^2}^{\textit{regularization}}$$

s.t.: $y\eta \geq 1-\xi$

$\xi \geq 0$

Support vector machines

$$\hat{\beta} = \sum_{i=1}^{N} \alpha_i y_i x_i$$

where $\alpha_i = 0$ for all x_i not on the margin

all x_i where $\alpha_i \neq 0$ are the support vectors

Same form as perceptron (with different semantics for α)

Support vectors

$$\hat{\beta} = \sum_{i=1}^{N} \alpha_i y_i x_i$$

- The support vectors are the small set of training data points that are most important for determining the decision boundary

Support vector machines

Predictions

$$\hat{y} = \hat{\beta}^\top x$$

$$\hat{y} = \sum_{i=1}^{N} \alpha_i y_i x_i^\top x$$

$$\hat{y} = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x)$$

Multiclass SVM

SVMs are inherently binary

One-versus-rest: K classifiers, one for each class versus all other classes

One-versus-one: K(K-1)/2 classifiers, one for each pair of classes

classification, so far

Genre classification

[TABLE1] TYPICAL FEATURES USED TO CHARACTERIZE MUSIC CONTENT.

TIMBRE	MELODY/HARMONY	RHYTHM
TEXTURE MODEL: MODEL OF FEATURES OVER TEXTURE WINDOW: 1) SIMPLE MODELING WITH LOW-ORDER STATISTICS 2) MODELING WITH AUTOREGRESSIVE MODEL 3) MODELING WITH DISTRIBUTION ESTIMATION ALGORITHMS (FOR EXAMPLE, EM ESTIMATION OF A GMM OF FRAMES)	PITCH FUNCTION: MEASURE OF THE ENERGY IN FUNCTION OF MUSIC NOTES 1) UNFOLDED FUNCTION: DESCRIBES PITCH CONTENT AND PITCH RANGE 2) FOLDED FUNCTION: DESCRIBES HARMONIC CONTENT	PERIODICITY FUNCTION: MEASURE OF THE PERIODICITIES OF FEATURES 1) TEMPO: PERIODICITIES TYPICALLY IN THE RANGE 0.3–1,5S (I.E., 200–40 BPM) 2) MUSICAL PATTERN: PERIODICITIES BETWEEN 2 AND 6 S (CORRESPONDING TO THE LENGTH OF ONE OR MORE MEASURE BAR)

[TABLE3] CONFUSION MATRIX FOR THE DATASET I AND FOR THE ALGORITHM SUBMITTED BY THE AUTHORS TO MIREX 2005.

TRUTH PREDICTION	AMBIENT	BLUES	CLASSIC	ELECTRONIC	ETHNIC	FOLK	JAZZ	NEW-AGE	PUNK	ROCK
AMBIENT	**52.94%**	0.00%	0.00%	7.32%	4.82%	0.00%	0.00%	26.47%	0.00%	5.95%
BLUES	0.00%	**76.47%**	0.00%	0.00%	0.00%	4.17%	0.00%	0.00%	0.00%	3.57%
CLASSIC	2.94%	0.00%	**100.00%**	0.00%	8.43%	0.00%	0.00%	0.00%	0.00%	0.00%
ELECTRONIC	5.88%	0.00%	0.00%	**53.66%**	6.02%	4.17%	4.55%	5.88%	0.00%	19.05%
ETHNIC	2.94%	0.00%	0.00%	7.32%	**59.04%**	12.50%	4.55%	20.59%	0.00%	0.00%
FOLK	0.00%	5.88%	0.00%	1.22%	3.61%	**62.50%**	0.00%	2.94%	0.00%	2.38%
JAZZ	0.00%	2.94%	0.00%	3.66%	6.02%	4.17%	**81.82%**	8.82%	0.00%	5.95%
NEW AGE	29.41%	0.00%	0.00%	4.88%	4.82%	8.33%	4.55%	**32.35%**	0.00%	5.95%
PUNK	0.00%	0.00%	0.00%	0.00%	0.00%	4.17%	0.00%	0.00%	**100.00%**	4.76%
ROCK	5.88%	14.71%	0.00%	21.95%	7.23%	0.00%	4.55%	2.94%	0.00%	**52.38%**

Midterm report

- 4 pages, citing 10 relevant sources
- Be sure to consider feedback!
- Data collection should be completed
- You should specify a validation strategy to be performed at the end
- Present initial experimental results

http://mybinder.org/repo/dbamman/dds