SPARSE KERNEL-BASED MODELS FOR SPEECH RECOGNITION

P. Karsmakers

Promotors:
Prof. Dr. Ir. J.A.K. Suykens (supervisor)
Prof. Dr. Ir. H. Van hamme

Public Ph.D. Defense
May 2010

Outline

- Introduction
  - Automatic Speech Recognition
  - Kernel methods
  - Sparse models
  - Motivation for Kernel Methods in Automatic Speech Recognition
  - Challenges and objectives
  - Main contributions
- Fixed-size Multi-class Kernel Logistic Regression
- Sparse Conjugate Directions Pursuit with Kernels
- Segment-based Phone Recognition using Kernel Methods
- Conclusions
- Future Work

1. Introduction: Automatic speech recognition

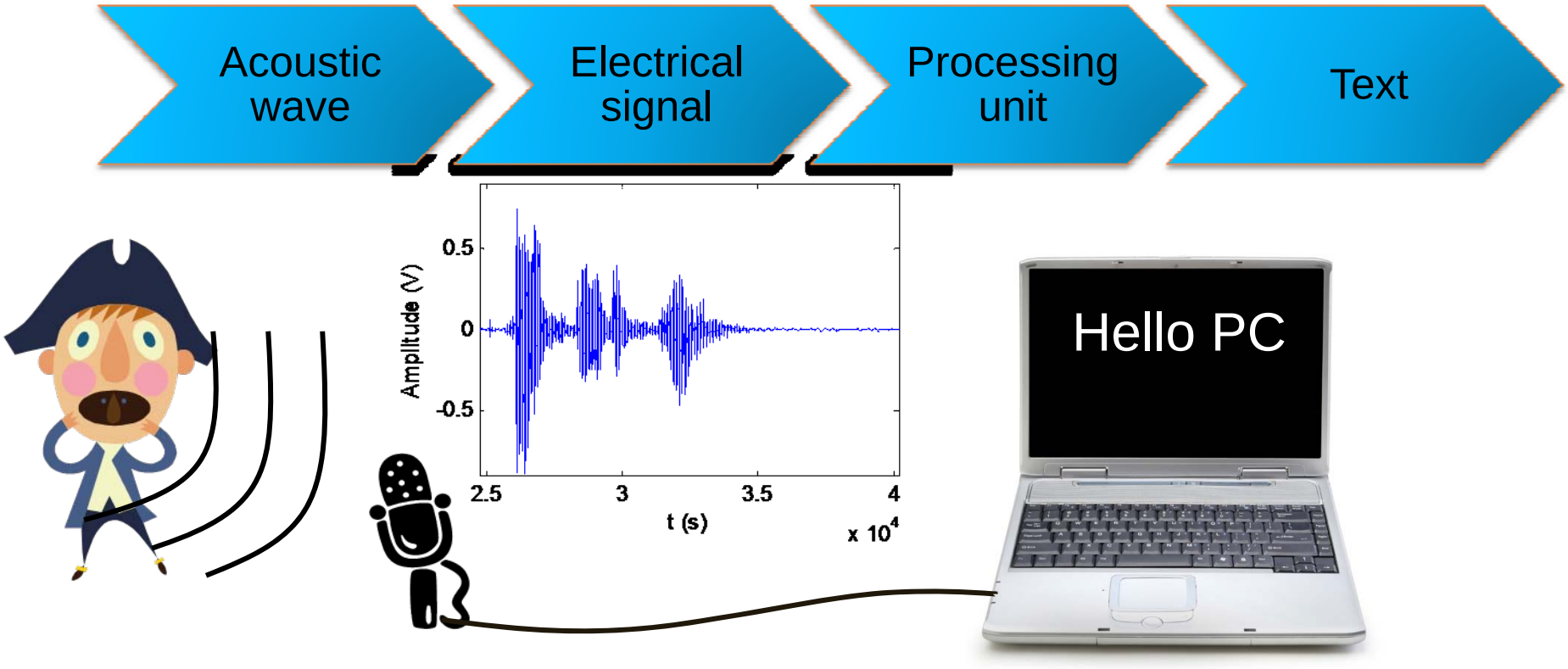
- Automatic Speech Recognition (ASR) = technology for converting an acoustic speech signal into a sequence of words by means of computer program.
- Many applications examples include:
  - Health:
    - Assistive technology: e.g. enable deaf people to understand spoken words, voice controlled home automation for people with mobility disabilities
  - Consumer electronics:
    - data entry and dictation
    - small mobile devices (e.g. smartphones) with voice dialing, voice controlled user interface
  - Military:
    - command and control e.g. in fighter aircraft with applications including setting radio frequencies, commanding an autopilot system

Different levels of complexity

- **Speech vocabulary:** the list of words which might be pronounced
- **Isolated or Continuous mode:** user clearly indicates word boundaries or not
- **Speaker dependent or independent:** system is developed for a single speaker or can be used by any speaker for a given language
- **Adverse environments:** mismatch between development and operational environment
- **Adaptivity:** ability of system to adjust to varying operating conditions

Experiments in this work are on continuous non-adaptive speaker-independent task in clean conditions

From acoustic wave to text



How to relate time-domain speech signals to text?

Acoustic model

- Relation too complex to be described by a set of rules, e.g.:

```
IF duration(x) < 1ms THEN y='hi'
IF max_amplitude(x) = 1 THEN y='peter'
…
```

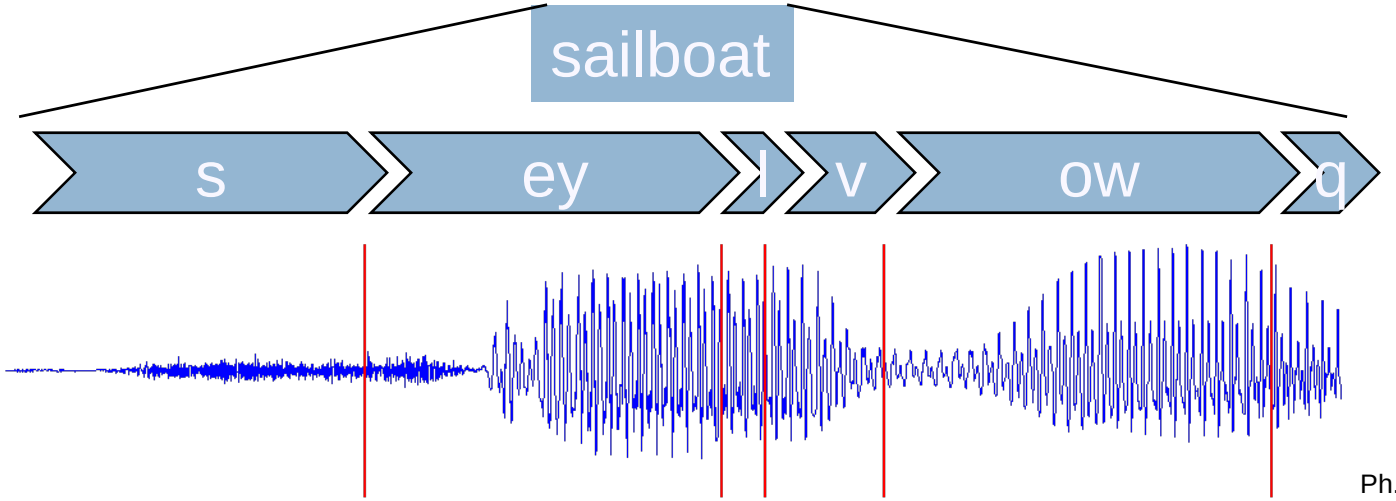
- Instead, let computer search for a (statistical) relation $y = f(x)$, acoustic model, based on annotated examples, e.g.:

x=Electrical signal	y=text
x_1= 	y_1='hello'
x_2= 	y_2='world'
x_3= 	y_3='steve'
…	…

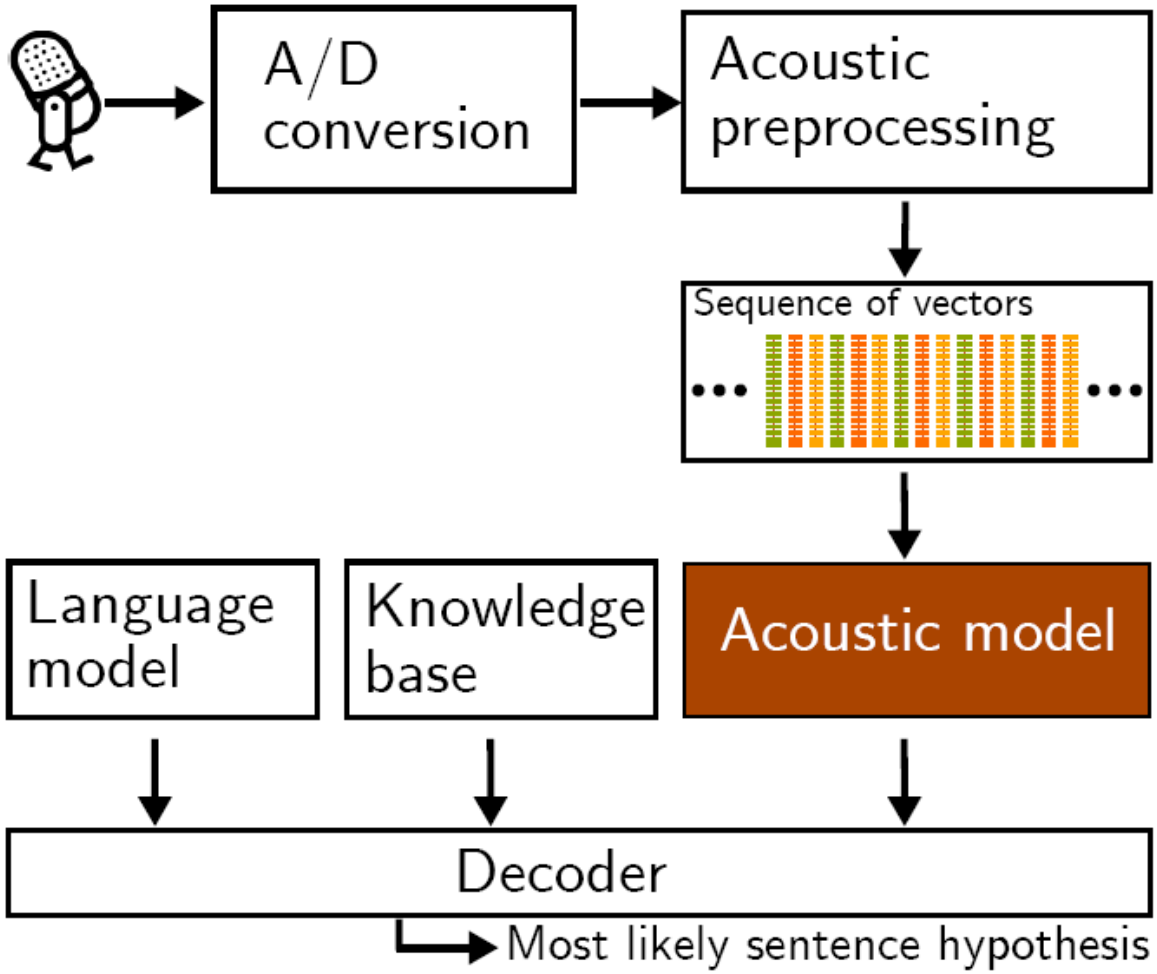
MACHINE LEARNING problem

Recognition of subwords

- Problem:
  - To obtain a good relation (model): each word needs a sufficient amount of examples. ±240.000 dutch words in Van Dale hence a lot of examples.
  - New words might appear
- Instead of modeling words directly use subwords (e.g. phone).
- Phone = smallest subword which has meaningful contrasts between utterances (about 50 phones)
- All words can be expressed by a combination of phones.

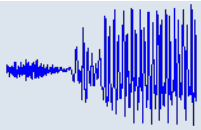
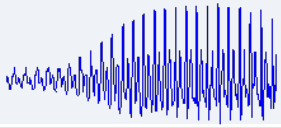
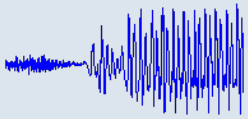


Classical speech recognizer architecture

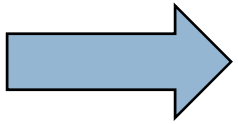


1. Introduction: Kernel-based models

- Learning an acoustic model = machine problem
- Kernel-based methods are specific family of machine learning methods
- Speech classification example

x=Electrical signal	y=text
x_1= 	y_1='ey'
x_2= 	y_2='ow'
x_3= 	y_3='ey'
…	…

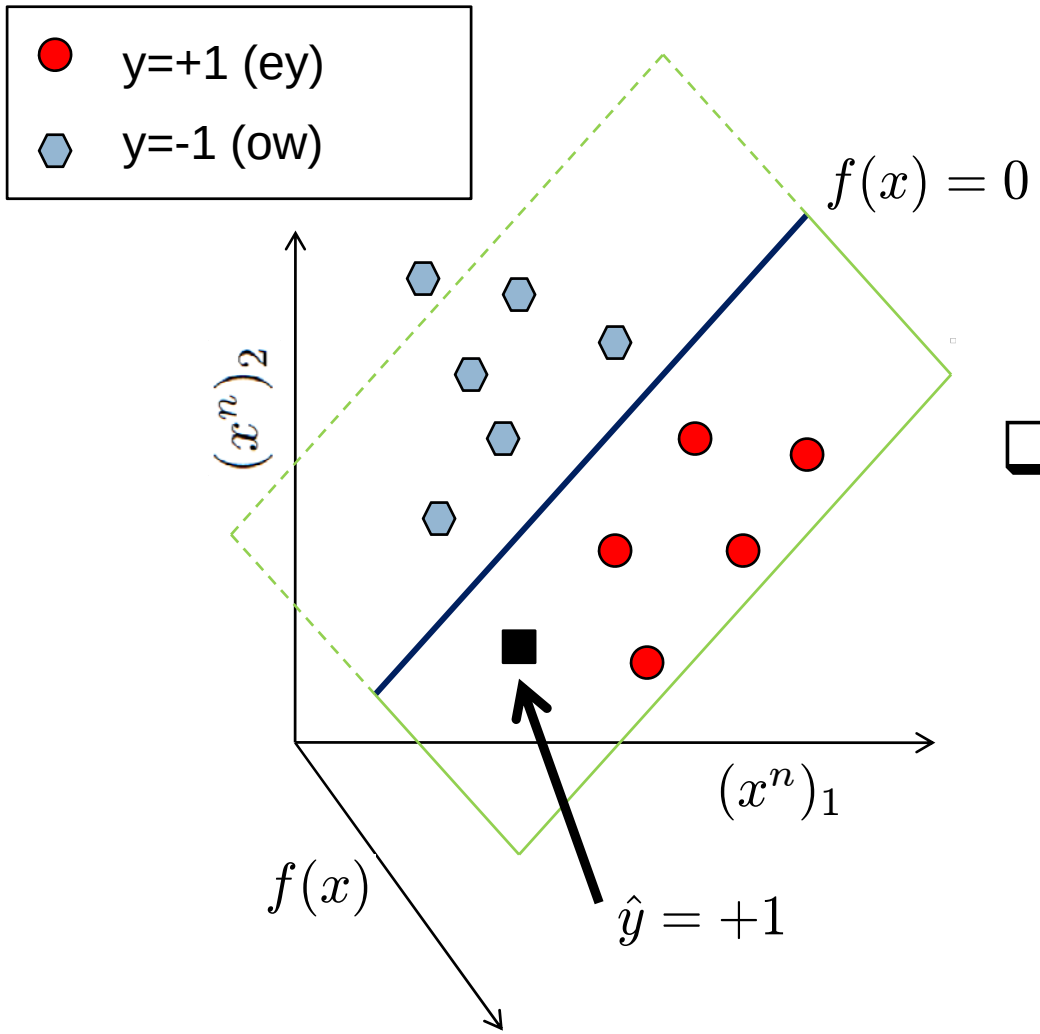
Different representation



x^n	y^n
$x^n_1=F(x_1)$	$y^n_1=+1$
$x^n_2=F(x_2)$	$y^n_2=-1$
$x^n_3=F(x_3)$	$y^n_3=+1$
…	…

- E.g. $F(x) : \mathbb{R}^M \rightarrow \mathbb{R}^2, F(x) = (mse(x), duration(x))^T$

Linear classification



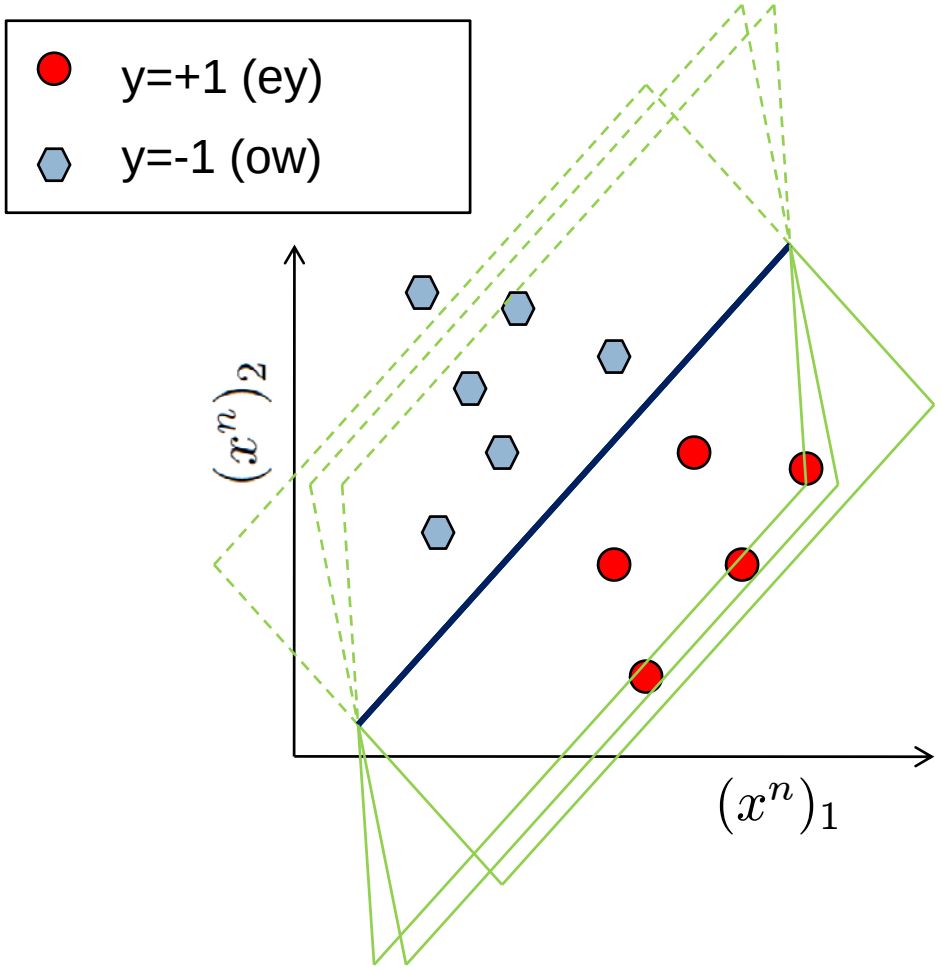
- Since classes are lineary separable, use linear model

$$f(x) = \sum_{i=1}^{2} w_i(x)_i + b$$

- Classify new point x* (■)

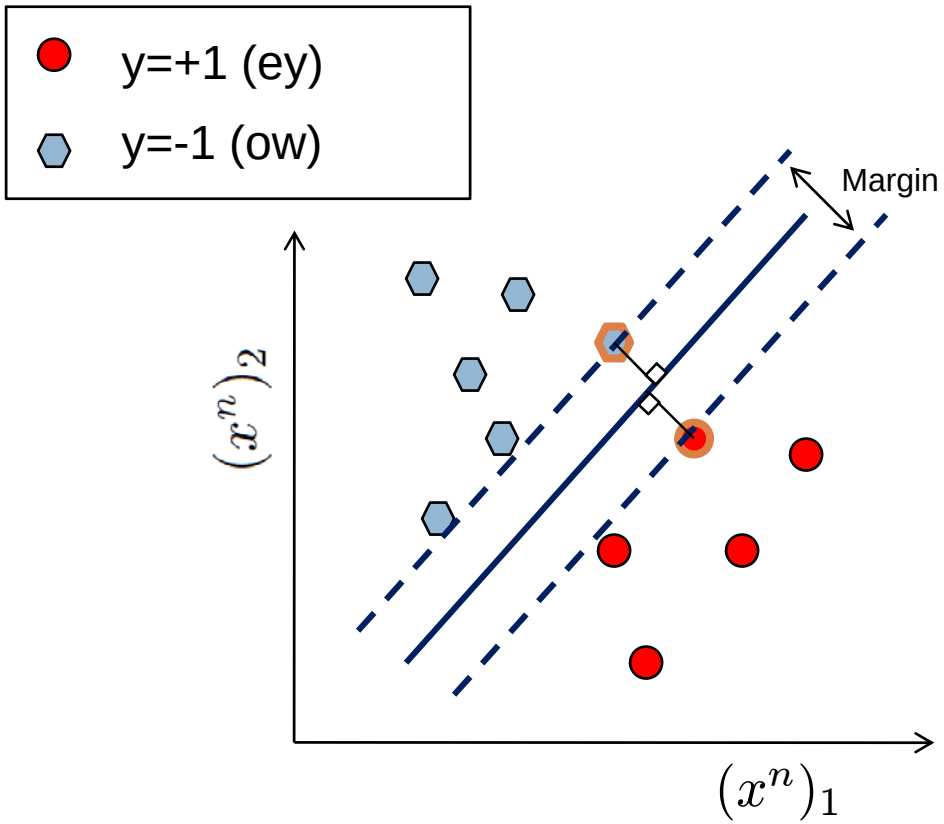
$$\hat{y} = \text{sign}(f(x))$$

Linear classification



- Which hyperplane to select?

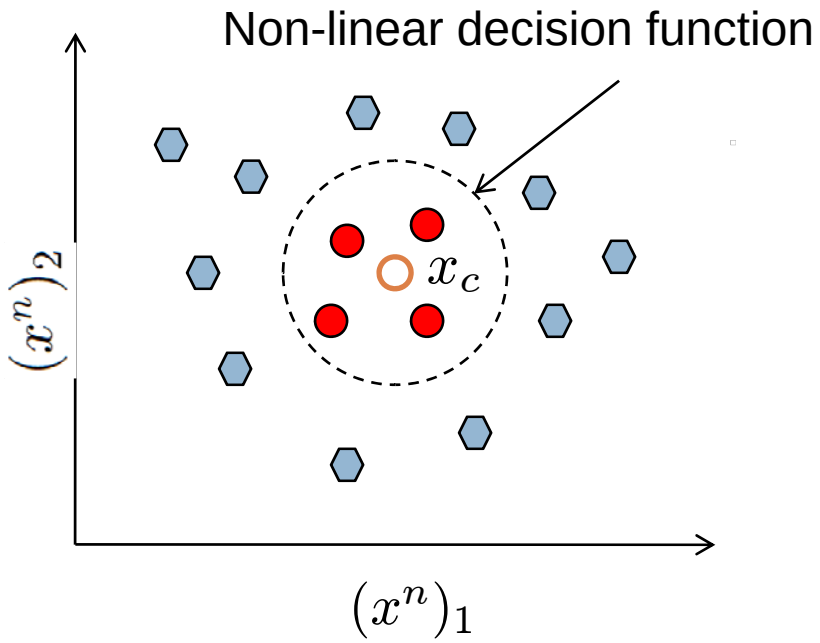
Linear classification



- Which hyperplane to select?
- Popular criterion → maximize the margin

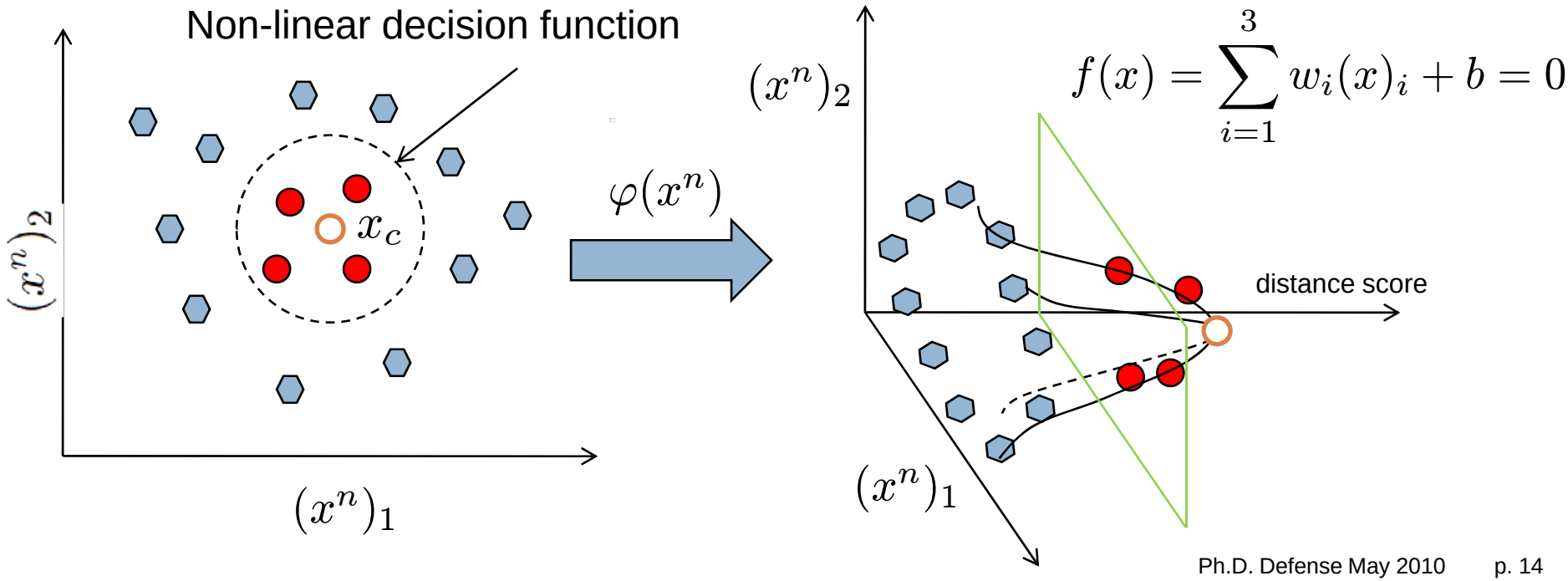
Non-linear classification

- What if data not linearly separable?



Non-linear classification

- What if data not linearly separable?
- Idea: first transform input data to higher dimensional space, then do linear classification
- E.g. use fixed mapping function $\varphi(x^n) = \left((x^n)_1, (x^n)_2, e^{\frac{-\sum_{i=1}^{2} ((x^n)_i - (x_c)_i)^2}{2\sigma^2}} \right)$



Kernel methods

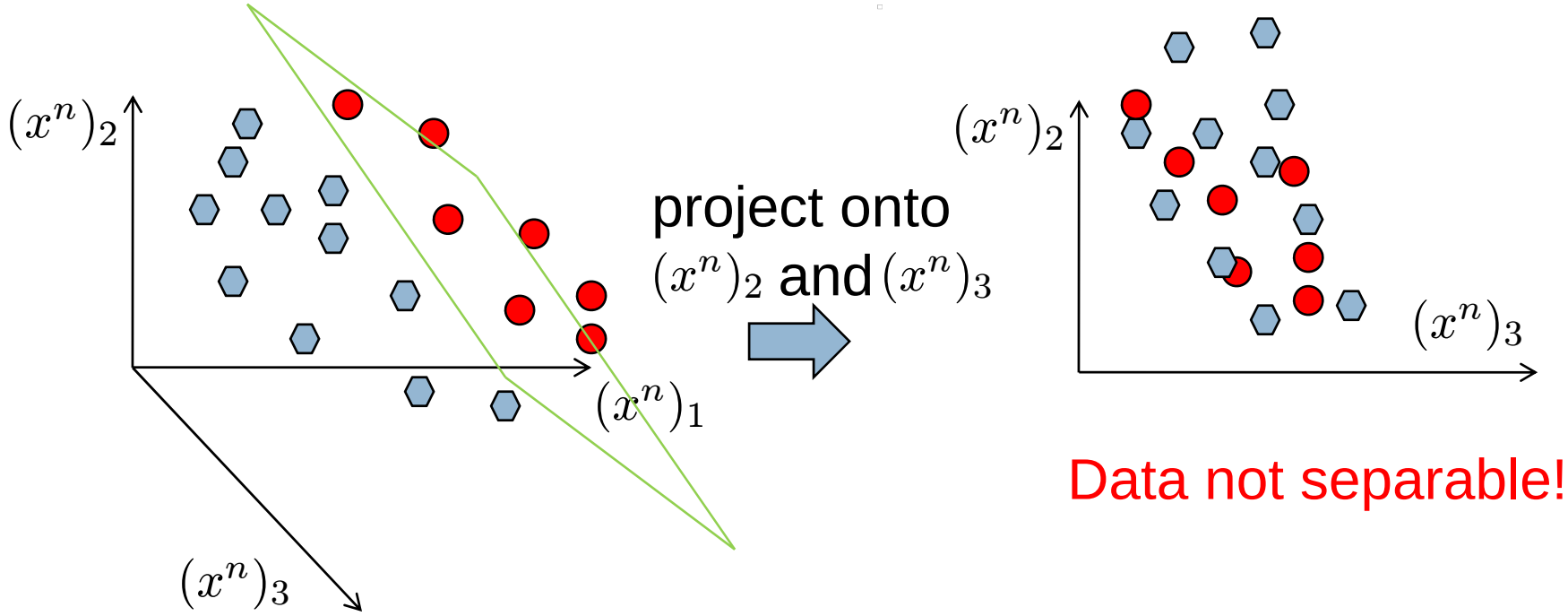
- Kernel methods do not explicitly define a feature map $\varphi(x)$ but implicitly define a feature map using a kernel function

  $$K(x, x') = \varphi(x)^T \varphi(x)$$

- Requires that learning problem is reformulated such that input data only appears in a dot-product
- Simply changing a single hyper-parameter results in different feature maps
- In higher dimensional space use well-understood algorithms to discover linear relations
- Support Vector Machines (SVMs) and Least-Squares Support Vector Machines (LS-SVMs) are popular implementations

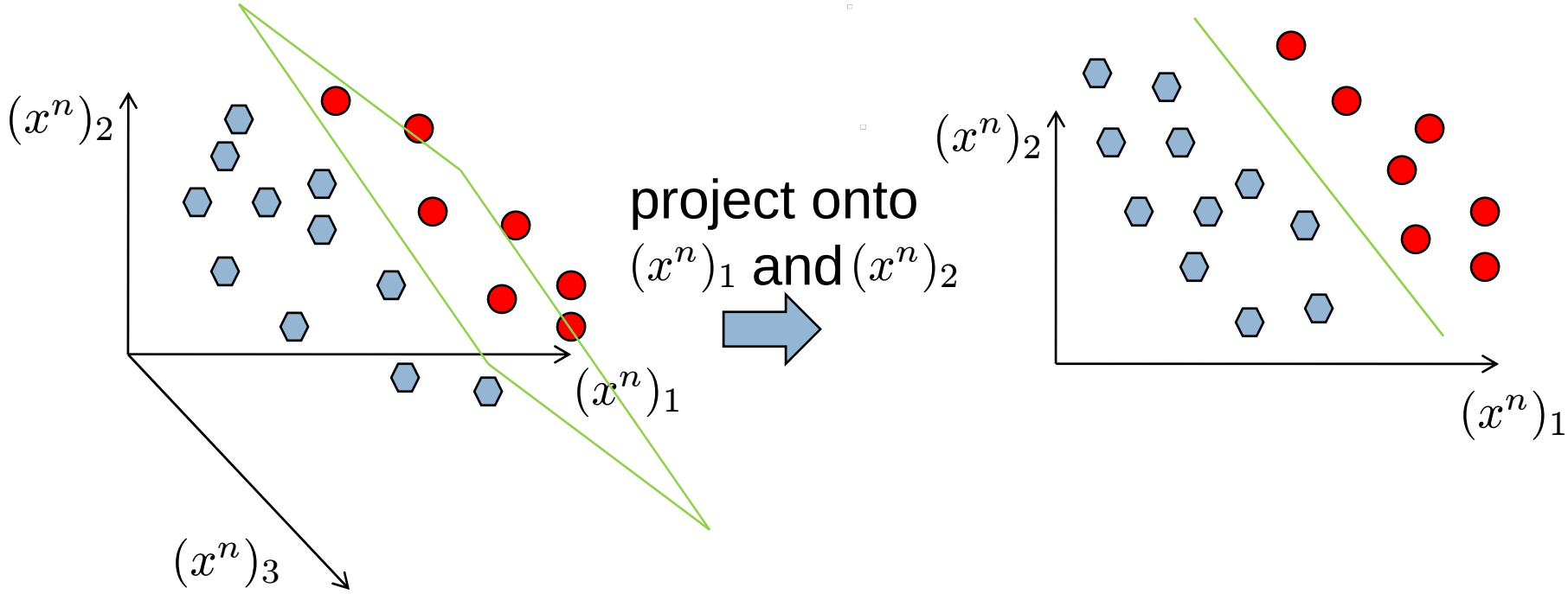
1. Introduction: Sparse models

- Keep the complexity of the model as low as possible.
- This might be obtained by setting a subset of the model parameters to zero → ignore corresponding data dimensions
- The more parameters are zero the sparser the model



1. Introduction: Sparse models

- Keep the complexity of the model as low as possible.
- This might be obtained by setting a subset of the model parameters to zero → ignore corresponding data dimensions
- The more parameters are zero the sparser the model



1. Introduction: Sparse models

- Without $(x^n)_3$ data is also separable
- No discriminative information in dimension $(x^n)_3$
- Set corresponding model parameter to zero
- $f(x) = (w)_1(x)_1 + (w)_2(x)_2 + 0.(x)_3 + b$, third dimension is ignored
- We have a sparse model!

1. Introduction: Motivation for Kernel Methods in Automatic Speech Recognition

- **ASR problem is far from solved**
- **Kernel methods might offer the following advantages:**
  - **Convexity:** Model parameters in ASR usually found using non-convex optimization → might give suboptimal results;
    Kernel methods such as SVM and LS-SVM are convex
  - **High dimensional spaces:** Kernel-based methods generalize well even in high dimensional spaces
  - **Customized kernels:** Different learning problems might be tackled using the same methodology and a different type of kernel.
  - **Success in other applications:** e.g. in bioinformatics, financial engineering, or time series prediction

1. Introduction: Challenges and Objectives

Requirements in ASR:

- Scalability: Training scales $O(N^2)$
- Sparseness: Sparser models give faster evaluation times
- Multi-class classification: Binary case typically extended with, mostly ad-hoc, approaches (e.g. one-versus-one, one-versus-all)
- Probabilistic outcomes: As a common interface in ASR a probabilistic language is normally used.
- Variable-length sequences: Mapping from variable length sequence to a fixed representation.

1. Introduction: Main contributions

- A. Large-scale Fixed-size Multi-class Kernel Logistic Regression (FS-MKLR)
  - P. Karsmakers, K. Pelckmans, H. Van hamme, J.A.K Suykens, "Large-Scale Kernel Logistic Regression for Segment-Based Phoneme Recognition",Internal Report 09-174, ESAT-SISTA, K.U.Leuven, Leuven, 2009, submitted for publication.
  - P. Karsmakers, K. Pelckmans, J. A. K. Suykens, "Multi-class kernel logistic regression: a fixed-size implementation," In Proc. of the international joint conference in neural networks (IJCNN),Orlando, Florida, U.S.A, pp. 1756-1761, 2007.
- B. Sparse Conjugate Directions Pursuit (SCDP)
  - P. Karsmakers, K. Pelckmans, K. De Brabanter, H. Van hamme, J.A.K Suykens, "Sparse Conjugate Directions Pursuit with Application to Fixed-size Kernel Models",Internal Report, ESAT-SISTA, K.U.Leuven, Leuven, 2010, submitted for publication.

1. Introduction: Main contributions

- C. Kernel-based phone classification and recognition using segmental features
  - P. Karsmakers, K. Pelckmans, H. Van hamme, J.A.K Suykens, "Large-Scale Kernel Logistic Regression for Segment-Based Phoneme Recognition",Internal Report 09-174, ESAT-SISTA, K.U.Leuven, Leuven, 2009, submitted for publication.
  - P. Karsmakers, K. Pelckmans, J. A. K. Suykens, H. Van hamme, "Fixed-Size Kernel Logistic Regression for Phoneme Classication,"In Proc. of INTERSPEECH, Antwerpen, Belgium, pp. 78-81, 2007.

2. Fixed-size Kernel Logistic Regression

- Focus on "kernelized" variant of Multi-class Logistic Regression (MKLR)
- Potential advantages:
  - Well-founded non-linear discriminative classifier
  - Yields a-posteriori probabilities of class membership based on maximum likelihood argument
  - Well described extension to multi-class case
- Potential disadvantages:
  - Scalability
  - Non-sparse models

Kernel Logistic Regression

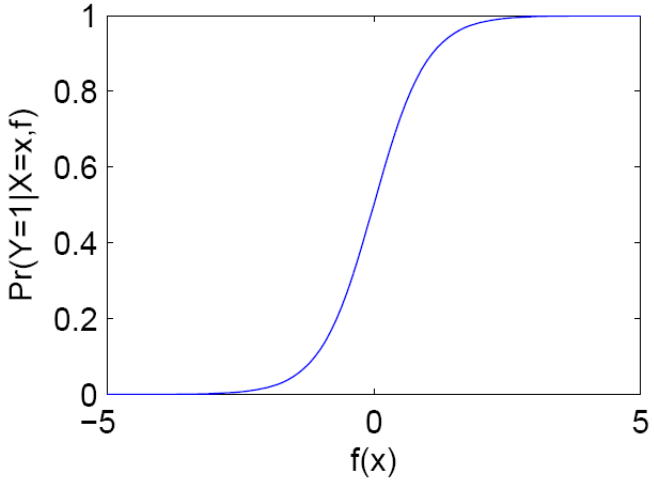
- Estimate a-posteriori probabilities using logistic function

Logistic model

$$\begin{cases} \Pr(\mathbf{Y} = -1 \mid \mathbf{X} = x; f) = \frac{\exp(f(x))}{1+\exp(f(x))} \\ \Pr(\mathbf{Y} = 1 \mid \mathbf{X} = x; f) = \frac{1}{1+\exp(f(x))}, \end{cases}$$

where $f(x) = w^T \varphi(x) + b$. Classification rule is

$$\underset{c \in \{-1,1\}}{\arg\max} \ \Pr(\mathbf{Y} = c | \mathbf{X} = x_*; f).$$



- Learning performed using a convex conditional maximum likelihood objective

Kernel Logistic Regression

- Equivalent dual formulation: using Newton's method, and standard LS-SVM approach the same model parameters can be obtained by iteratively solving

Dual formulation at iteration *k*

$$\left(\Omega + \frac{1}{\nu} G^{(k)^{-1}}\right) \alpha^{(k)} = \frac{1}{\nu} q^{(k)} + \Omega \alpha^{(k-1)},$$

where $\Omega_{ij} = \varphi(x_i)^T \varphi(x_j) = K(x_i, x_j)$, $G^{(k)}$ a weight matrix, and $q^{(k)}$ a regressor vector.

- Mapped input vectors only appear in dot-products
- Kernel functions can be used → non-linear classification

Fixed-size Kernel Logistic Regression

Fixed-Size Multi-class KLR (FS-MKLR) is proposed:

- Use all-at-once multi-class logistic model
- Explicit approximation of the nonlinear mapping $\hat{\varphi} : \mathbb{R} \rightarrow \mathbb{R}^M$ using Nyström.
- Based on subsample (set of Prototype Vectors (PVs)) of training set, selected using k-center clustering.
- Solve primal problem using customized Newton-trust region optimization for multi-class classification.

Advantages to classical MKLR

- Scalable to large-scale data sets (N > 50,000)
- Sparse models with easily manageable model size M

Selected experiments: Active PV selection methods

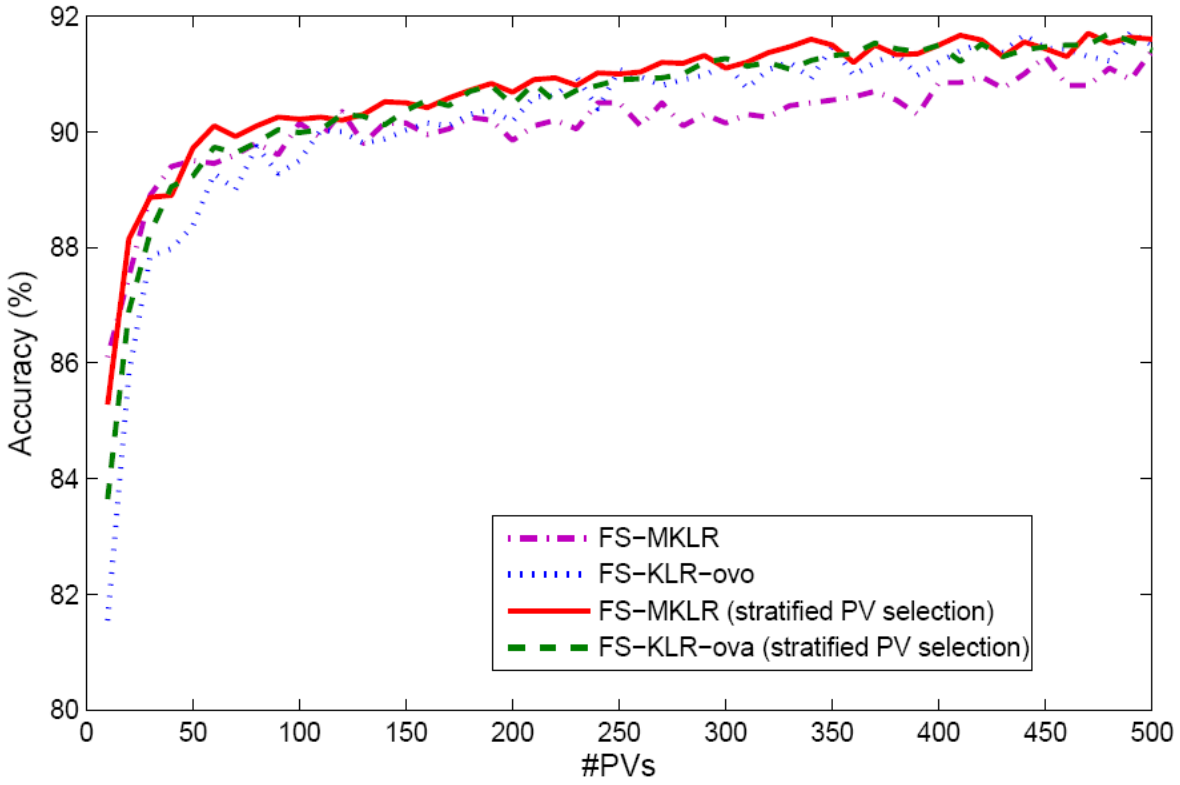
- PV selection important to approximate feature map
- Compare 3 different active PV selection methods on 11 benchmark data sets

	k-center		renyi		random	
	err	log likelihood	err	log likelihood	err	log likelihood
banana	**10.63**(±**0.5**)	−0.2463(±0.0101)	10.72(±0.5)	−0.2501(±0.0128)	10.68(±0.52)	−0.2501(±0.0114)
breast-cancer	27.52(±4.71)	−0.5426(±0.0485)	**27.42**(±**4.52**)	−0.5436(±0.047)	27.94(±4.53)	−0.5473(±0.0492)
diabetis	**23.23**(±**1.83**)	−0.4841(±0.0186)	23.3(±1.84)	−0.4856(±0.0184)	23.6(±1.68)	−0.4823(±0.0177)
flare-solar	33.67(±1.65)	−0.5725(±0.0139)	**33.62**(±**1.61**)	−0.573(±0.0136)	34.02(±1.88)	−0.5731(±0.0166)
heart	**16.7**(±**3.55**)	−0.4071(±0.0582)	18.33(±4.16)	−0.4295(±0.0648)	17.75(±3.49)	−0.4167(±0.0548)
ringnorm	**1.85**(±**0.28**)	−0.0575(±0.0156)	2.29(±0.34)	−0.0654(±0.009)	1.91(±0.3)	−0.0569(±0.0089)
splice	**16.61**(±**1.03**)	−0.3743(±0.0137)	17.78(±1.31)	−0.3926(±0.0183)	17.3(±0.92)	−0.3821(±0.0154)
thyroid	**4.95**(±**2.23**)	−0.1388(±0.1029)	5.67(±2.72)	−0.17(±0.0954)	5.36(±2.63)	−0.1482(±0.0973)
titanic	22.71(±0.98)	−0.5336(±0.039)	**22.62**(±**0.86**)	−0.6377(±0.8776)	23.13(±1.43)	−0.5727(±0.1812)
twonorm	**2.79**(±**0.32**)	−0.0769(±0.0072)	2.94(±0.33)	−0.0808(±0.0126)	2.99(±0.31)	−0.0805(±0.0082)
waveform	**10.13**(±**0.56**)	−0.224(±0.0105)	10.93(±0.47)	−0.2402(±0.0084)	10.92(±0.53)	−0.2402(±0.0105)

- On average k-center clustering gave the best results

Selected experiments: Sparsity of multi-class schemes

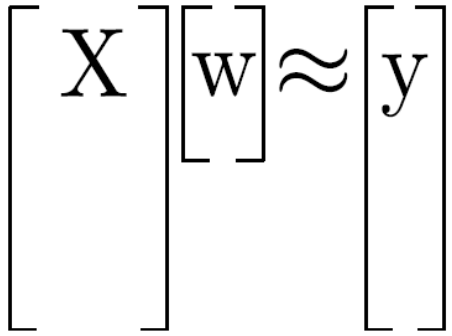
- Compared to combined binary classifiers, all-at-once multi-class approach (with stratified PV selection) is preferred.
- `satimage` data set (#classes = 6; #examples = 4,435; #dimensions = 36).



Main conclusions

- All-at-once FS-MKLR gave sparsest, and fastest models compared to one-versus-one coding scheme, while having similar or better accuracies.
- FS-MKLR models are far much sparser than SVM while obtaining comparable accuracy.
- Compared to its alternatives k-center clustering with outlier removal PV selection is preferred for KLR (stratified selection for multi-class).

3. Sparse Conjugate Directions Pursuit



- Suppose a set of *N* equations, and *D* unknowns.
- If *N*>*D* then over-determined linear system
- In general an over-determined system has no solution, therefore choose solution according to some optimality criterion, e.g.

Objective

$$w^{0,*} = \arg\min_{w} \|w\|_0 \quad \text{such that} \quad \|Xw - y\|_2^2 \leq \delta,$$

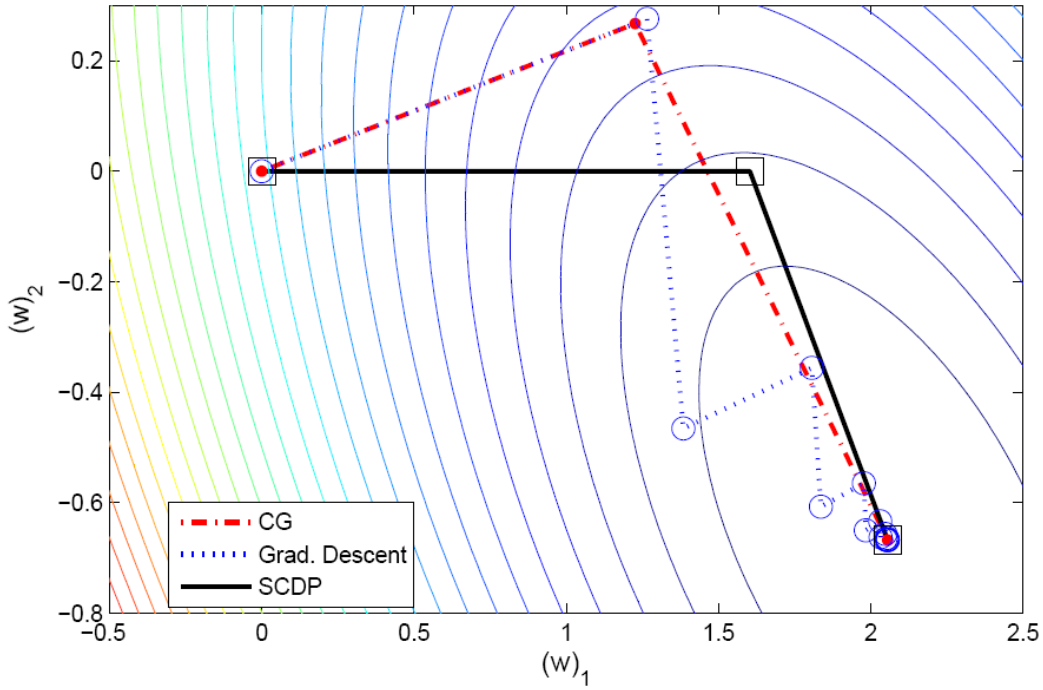
where $\delta > 0$ is an appropriate error tolerance.

Motivations for L_0-norm

- Estimation problems: sparse coefficients → feature selection
- Machine learning: sparse predictor rules → improved generalization
- Sparse solution leads to computation and memory-efficient model evaluations
- Sparse solution might be exploited when designing scalable algorithms

Heuristically solve previous objective by adapting Conjugate Gradient (CG):

- Iteratively construct a sparse conjugate basis $\|w^{(k)}\|_0 = k$, starting with $w^{(1)} = 0_D$.
- Globally optimal, local optimization associated to each conjugate basis vector.
- Sequence of conjugate basis vectors is such that a small number of iterations suffice, leading to a sparse solution.



❑ We call this algorithm Sparse Conjugate Directions Pursuit (SCDP).

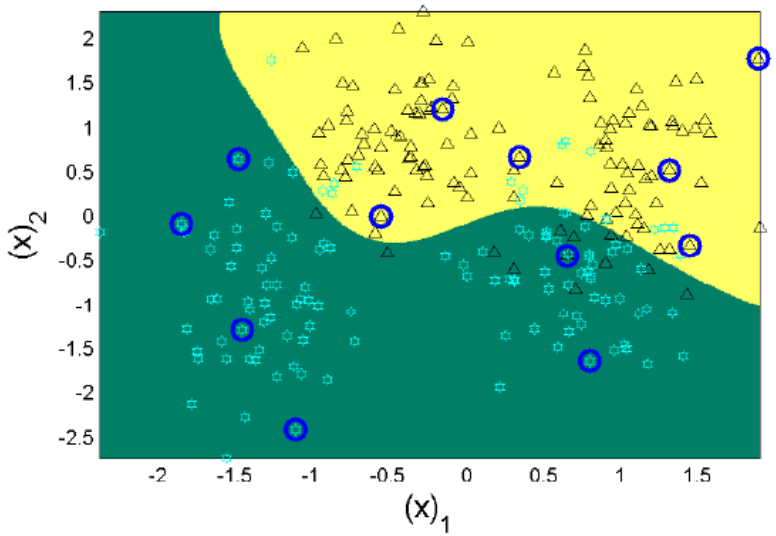
Application: using SCDP for sparse reduced LS-SVMs

Based on SCDP a new kernel-based learning method is derived within the LS-SVM setting:

- Fixed-size LS-SVM approximately solves the primal LS-SVM problem, based on *M* PVs (selected beforehand) Nyström approximation
- Leads to an over-determined linear system of size $N >> M$.
- SCDP is then applied to get sparser models $||w||_0 < M$ (SCDP-FSLSSVM).

Selected experiments

- Decision boundary and final PV positions when using SCDP-FSLSSVM on `ripley` benchmark



- SCDP-FSLSSVM much sparser than SVM while having similar performance.

Data set	SCDP-FSLSSVM		SVM	
	err (%)	# PV	err (%)	# SV
adult	$14.55(\pm 0.34)$	$166(\pm 41.6)$	$14.53(\pm 0.14)$	$10,494(\pm 212.4.0)$
gamma telescope	$13.90(\pm 0.27)$	$778(\pm 58.9)$	$13.3(\pm 0.12)$	$4,211(\pm 112.6)$

Main conclusions

- SCDP-FSLSSVM and LS-SVM have similar accuracies, SCDP-FSLSSVM gives sparse models and has faster training.
- Similar prediction accuracies as those of SVM were obtained, while SCDP usually produces much sparser models.
- Compared to SVM and LS-SVM, SCDP-FSLSSVM is not a convex learning method. However, for a given set of PVs (possibly all training data) and $w^{(1)} = 0_D$ SCDP-FSLSSVM gives a unique solution.
- k-center clustering as a preprocessing step to select initial PVs, speed-up training.

4. Segment-based Phone Recognition

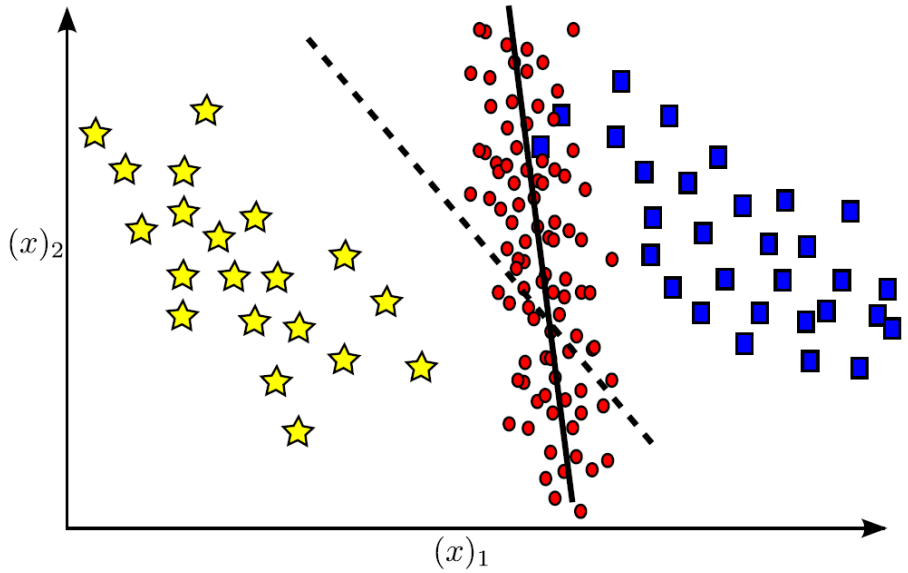
- Recall our feature set $F(x) : \mathbb{R}^M \rightarrow \mathbb{R}^2, F(x) = (mse(x), duration(x))^T$
- In practice more and better features needed
- State-of-the-art uses features computed for very small speech parts
- In our setup we use features computed for larger parts, i.e. phone segments
- Focus on recognizing phone sequences, forms a basis for a good word recognition

Motivation: Kernel Models in Segment-based Approach

- State-of-the-art in ASR based on Hidden Markov Models (HMM)
- Impressive HMMs but modeling limitations, e.g. duration- and trajectory modeling.
- Segment-based setup might overcome some.
- Segments span larger context windows, possibly more features are needed.
- Kernel-based methods generalize well even in high dimensional spaces (many features are used)

Inference with a universum

- **add**



Selected experiments

- On timit speech corpus.
- Segment-based classification and recognition. Segment (unit)= phone
- Training set size: 142, 910 train vectors and 51, 681 test vectors with 181 dimensions

Phone Classification

- **ADD**

Phone Recognition

- **ADD**

Main conclusions

- Phone classification:
  - Kernel based alternatives outperformed a state-of-the-art HMM classifier
  - FS-MKLR, but also SCDP-FSLSSVM and SVM (only indirectly estimate a-posteriori probabilities) match the Bayes probability fairly well.
- Phone recognition:
  - Although room left for improvement, without LM segment-based approach has state-of-the-art accuracy.
  - Universum data improved the final PER without increasing the number of parameters of the phone model.
  - However, less gain when using LMs as is the case for the HMM recognizer.

5. General Conclusions

- Previously specified requirements were tackled as follows:
  - Scalability: Two practical, and scalable kernel-based algorithms: FS-MKLR and SCDP-FSLSSVM. Trade-off between model accuracy and training (and model) complexity, directly controlled by the user.
  - Multi-class classification: All-at-once FS-MKLR is preferred compared to binary coupled variants in terms of classification accuracy and model sparsity.
  - Sparseness: (i) Tuned SVM model not really sparse;(ii) One-versus-one coding scheme additionally increases cardinality; (iii) Proposed methods produce significantly sparser models while having comparable accuracies.

5. General Conclusions

- Probabilistic interpretation: All considered methods yield, either direct or indirect, probabilistic outcomes, both empirically give adequate results.

- Variable-length sequence: A simple and fast mapping to fixed-length vectors was used.

- We successfully integrated our new kernel models in a segment-based speech recognition system and compared to a state-of-the-art ASR system

6. Future Work

- Segmentation model: use a more sophisticated model, possibly using other types of features

- Design of a customized kernel: use other positive semi-definite kernels, e.g. sequence kernels → score pairs of variable-length segments directly

- SCDP-FSMKLR: apply SCDP on FS-MKLR method.