Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

April 5, 2011

Today:

- Latent Dirichlet Allocation
  - topic models
- Social network analysis based on latent probabilistic models
- Kernel regression

Readings:

- Kernels: Bishop Ch. 6.1

optional:

- Bishop Ch 6.2, 6.3
- "Kernel Methods for Pattern Analysis", Shawe-Taylor & Cristianini, Chapter 2

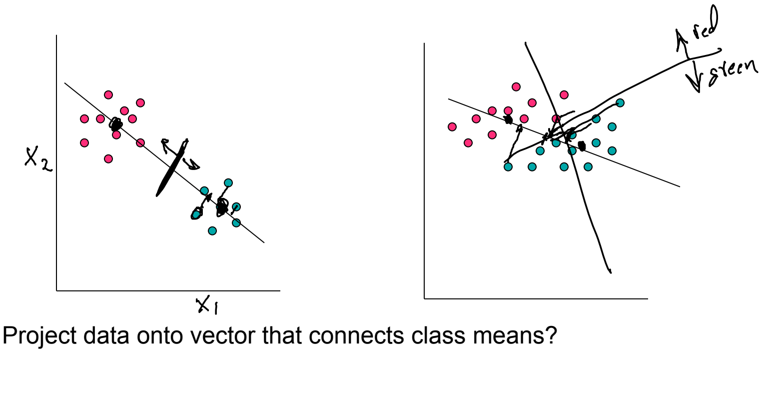
Supervised Dimensionality Reduction

Supervised Dimensionality Reduction

- Neural nets: learn hidden layer representation, designed to optimize network prediction accuracy
- PCA: unsupervised, minimize reconstruction error
  - but sometimes people use PCA to re-represent original data before classification (to reduce dimension, to reduce overfitting)
- Fisher Linear Discriminant
  - like PCA, learns a *linear* projection of the data
  - but supervised: it uses labels to choose projection

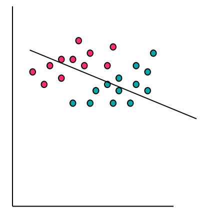
Fisher Linear Discriminant

- A method for projecting data into lower dimension to hopefully improve classification
- We'll consider 2-class case



Project data onto vector that connects class means?

Fisher Linear Discriminant



Project data onto one dimension, to help classification

$$y = \mathbf{w}^T \mathbf{x}$$

Define class means: $\mathbf{m}_i = \frac{1}{N_i} \sum_{n \in C_i} \mathbf{x}^n$

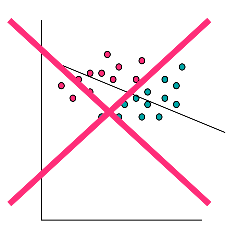
Could choose w according to: $\arg\max_{\mathbf{w}} \mathbf{w}^T (\mathbf{m_2} - \mathbf{m_1})$

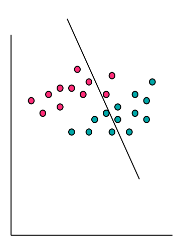
Instead, Fisher Linear Discriminant chooses: $\boxed{\arg\max_{\mathbf{w}} \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}}$

$$m_i \equiv \mathbf{w}^T \mathbf{m}_i \qquad s_i^2 \equiv \sum_{n \in C_i} (x^n - m_i)^2$$

Summary: Fisher Linear Discriminant

- Choose n-1 dimension projection for n-class classification problem
- Use within-class covariances to determine the projection
- Minimizes a different error function (the projected within-class variances)





Example topics induced from a large collection of text

DISEASE	WATER	MIND	STORY	FIELD	SCIENCE	BALL	JOB
BACTERIA	FISH	WORLD	STORIES	MAGNETIC	STUDY	GAME	WORK
DISEASES	SEA	DREAM	TELL	MAGNET	SCIENTISTS	TEAM	JOBS
GERMS	SWIM	DREAMS	CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
FEVER	SWIMMING	THOUGHT	CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CAUSE	POOL	IMAGINATION	AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
CAUSED	LIKE	MOMENT	READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
SPREAD	SHELL	THOUGHTS	TOLD	POLES	CHEMISTRY	FIELD	WORKING
VIRUSES	SHARK	OWN	SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
INFECTION	TANK	REAL	TALES	COMPASS	MANY	BASKETBALL	SKILLS
VIRUS	SHELLS	LIFE	PLOT	LINES	MATHEMATICS	COACH	CAREERS
MICROORGANISMS	SHARKS	IMAGINE	TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
PERSON	DIVING	SENSE	SHORT	ELECTRIC	FIELD	PLAYING	FIND
INFECTIOUS	DOLPHINS	CONSCIOUSNESS	FICTION	DIRECTION	PHYSICS	HIT	POSITION
COMMON	SWAM	STRANGE	ACTION	FORCE	LABORATORY	TENNIS	FIELD
CAUSING	LONG	FEELING	TRUE	MAGNETS	STUDIES	TEAMS	OCCUPATIONS
SMALLPOX	SEAL	WHOLE	EVENTS	BE	WORLD	GAMES	REQUIRE
BODY	DIVE	BEING	TELLS	MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
INFECTIONS	DOLPHIN	MIGHT	TALE	POLE	STUDYING	BAT	EARN
CERTAIN	UNDERWATER	HOPE	NOVEL	INDUCED	SCIENCES	TERRY	ABLE

[Tennenbaum et al]

What about Probabilistic Approaches?

Supervised?

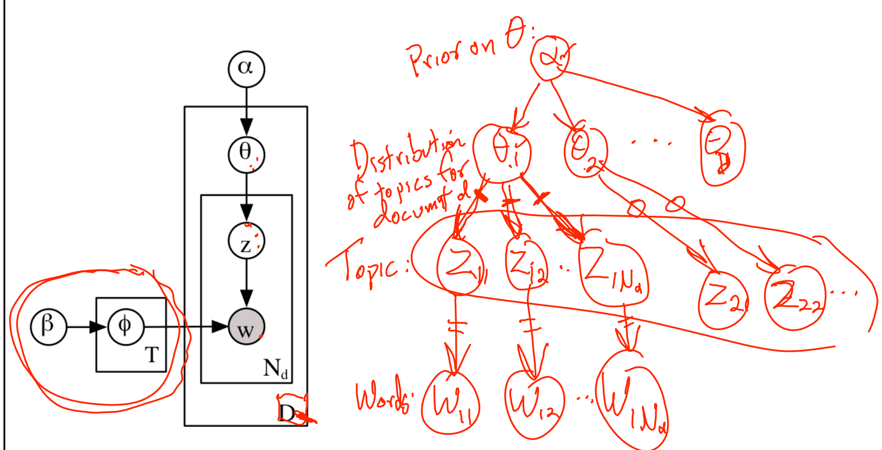
Unsupervised?

Example topics induced from a large collection of text

DISEASE	WATER	MIND	STORY	FIELD	SCIENCE	BALL	JOB
BACTERIA	FISH	WORLD	STORIES	MAGNETIC	STUDY	GAME	WORK
DISEASES	SEA	DREAM	TELL	MAGNET	SCIENTISTS	TEAM	JOBS
GERMS	SWIM	DREAMS	CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
FEVER	SWIMMING	THOUGHT	CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CAUSE	POOL	IMAGINATION	AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
CAUSED	LIKE	MOMENT	READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
SPREAD	SHELL	THOUGHTS	TOLD	POLES	CHEMISTRY	FIELD	WORKING
VIRUSES	SHARK	OWN	SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
INFECTION	TANK	REAL	TALES	COMPASS	MANY	BASKETBALL	SKILLS
VIRUS	SHELLS	LIFE	PLOT	LINES	MATHEMATICS	COACH	CAREERS
MICROORGANISMS	SHARKS	IMAGINE	TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
PERSON	DIVING	SENSE	SHORT	ELECTRIC	FIELD	PLAYING	FIND
INFECTIOUS	DOLPHINS	CONSCIOUSNESS	FICTION	DIRECTION	PHYSICS	HIT	POSITION
COMMON	SWAM	STRANGE	ACTION	FORCE	LABORATORY	TENNIS	FIELD
CAUSING	LONG	FEELING	TRUE	MAGNETS	STUDIES	TEAMS	OCCUPATIONS
SMALLPOX	SEAL	WHOLE	EVENTS	BE	WORLD	GAMES	REQUIRE
BODY	DIVE	BEING	TELLS	MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
INFECTIONS	DOLPHIN	MIGHT	TALE	POLE	STUDYING	BAT	EARN
CERTAIN	UNDERWATER	HOPE	NOVEL	INDUCED	SCIENCES	TERRY	ABLE

[Tennenbaum et al]

Plate Notation



Latent Dirichlet Allocation Model

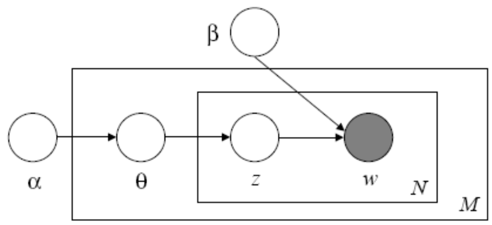


Figure 1: Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

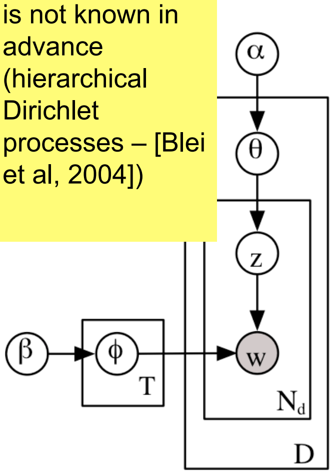
where $p(z_n | \theta)$ is simply θ_i for the unique i such that $z_n^i = 1$. Integrating over θ and summing over z, we obtain the marginal distribution of a document:

$$p(\mathbf{w} | \alpha, \beta) = \int p(\theta | \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \beta) \right) d\theta. \qquad (3)$$

ustering words into topics with Latent Dirichlet Allocation

[Blei, Ng, Jordan 2003]

Also extended to case where number of topics is not known in advance (hierarchical Dirichlet processes – [Blei et al, 2004])



Probabilistic model for document set:

For each of the D documents:

1. Pick a $\theta_d \sim P(\theta|\alpha)$ to define $P(z|\theta_d)$
2. For each of the N_d words w
    - Pick topic $z_n \sim P(z \mid \theta_d)$
    - Pick word $w_n \sim P(w \mid z_n, \phi)$

Training this model defines topics (i.e., ϕ which defines P(W|Z))

Example topics induced from a large collection of text

Significance:

• Learned topics reveal implicit semantic categories of words within the documents

• In many cases, we can represent documents with 10^2 topics instead of 10^5 words

• Especially important for short documents (e.g., emails). Topics overlap when words don't !

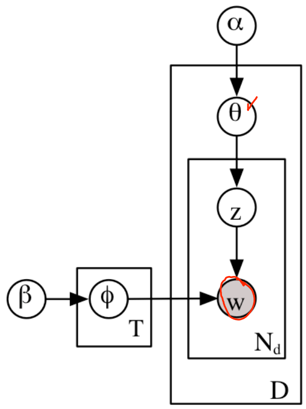
FIELD	SCIENCE	BALL	JOB
MAGNETIC	STUDY	GAME	WORK
MAGNET	SCIENTISTS	TEAM	JOBS
WIRE	SCIENTIFIC	FOOTBALL	CAREER
NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CURRENT	WORK	PLAYERS	EMPLOYMENT
COIL	RESEARCH	PLAY	OPPORTUNITIES
POLES	CHEMISTRY	**FIELD**	WORKING
IRON	TECHNOLOGY	PLAYER	TRAINING
COMPASS	MANY	BASKETBALL	SKILLS
LINES	MATHEMATICS	COACH	CAREERS
CORE	BIOLOGY	PLAYED	POSITIONS
ELECTRIC	**FIELD**	PLAYING	FIND
DIRECTION	PHYSICS	HIT	POSITION
FORCE	LABORATORY	TENNIS	**FIELD**
MAGNETS	STUDIES	TEAMS	OCCUPATIONS
BE	WORLD	GAMES	REQUIRE
MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
POLE	STUDYING	BAT	EARN
INDUCED	SCIENCES	TERRY	ABLE

[Tennenbaum et al]

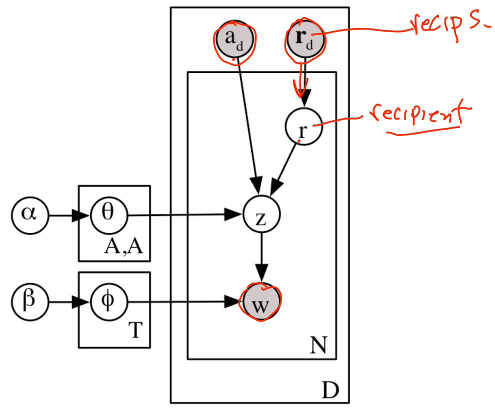
Analyzing topic distributions in email

Author-Recipient-Topic model for Email

Latent Dirichlet Allocation
(LDA)
[Blei, Ng, Jordan, 2003]



Author-Recipient Topic
(ART)
[McCallum, Corrada, Wang, 2005]



Enron Email Corpus

- 250k email messages
- 23k people

```
Date: Wed, 11 Apr 2001 06:56:00 -0700 (PDT)
From: debra.perlingiere@enron.com
To: steve.hooser@enron.com
Subject: Enron/TransAltaContract dated Jan 1, 2001

Please see below. Katalin Kiss of TransAlta has requested an
electronic copy of our final draft?  Are you OK with this?  If
so, the only version I have is the original draft without
revisions.

DP

Debra Perlingiere
Enron North America Corp.
Legal Department
1400 Smith Street, EB 3885
Houston, Texas 77002
dperlin@enron.com
```

Topics, and prominent sender/receivers discovered by ART

[McCallum et al, 2005]

Top words within topic :

Top author-recipients exhibiting this topic

Topic 17 "Document Review"		Topic 27 "Time Scheduling"		Topic 45 "Sports Pool"	
attached	0.0742	day	0.0419	game	0.0170
agreement	0.0493	friday	0.0418	draft	0.0156
review	0.0340	morning	0.0369	week	0.0135
questions	0.0257	monday	0.0282	team	0.0135
draft	0.0245	office	0.0282	eric	0.0130
letter	0.0239	wednesday	0.0267	make	0.0125
comments	0.0207	tuesday	0.0261	free	0.0107
copy	0.0165	time	0.0218	year	0.0106
revised	0.0161	good	0.0214	pick	0.0097
document	0.0156	thursday	0.0191	phillip	0.0095
G.Nemec B.Tycholiz	0.0737	J.Dasovich R.Shapiro	0.0340	E.Bass M.Lenhart	0.3050
G.Nemec M.Whitt	0.0551	J.Dasovich J.Steffes	0.0289	E.Bass P.Love	0.0780
B.Tycholiz G.Nemec	0.0325	C.Clair M.Taylor	0.0175	M.Motley M.Grigsby	0.0522

Topics, and prominent sender/receivers discovered by ART

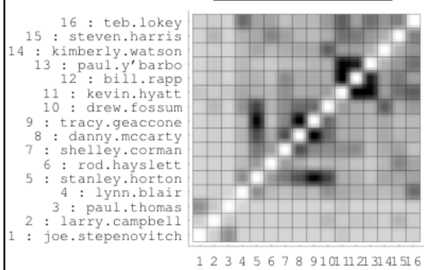
Topic 34 "Operations"		Topic 37 "Power Market"		Topic 41 "Government Relations"		Topic 42 "Wireless"	
operations	0.0321	market	0.0567	state	0.0404	blackberry	0.0726
team	0.0234	power	0.0563	california	0.0367	net	0.0557
office	0.0173	price	0.0280	power	0.0337	www	0.0409
list	0.0144	system	0.0206	energy	0.0239	website	0.0375
bob	0.0129	prices	0.0182	electricity	0.0203	report	0.0373
open	0.0126	high	0.0124	davis	0.0183	wireless	0.0364
meeting	0.0107	based	0.0120	utilities	0.0158	handheld	0.0362
gas	0.0107	buy	0.0117	commission	0.0136	stan	0.0282
business	0.0106	customers	0.0110	governor	0.0132	fyi	0.0271
houston	0.0099	costs	0.0106	prices	0.0089	named	0.0260
S.Beck L.Kitchen	0.2158	J.Dasovich J.Steffes	0.1231	J.Dasovich R.Shapiro	0.3338	R.Haylett T.Geaccone	0.1432
S.Beck J.Lavorato	0.0826	J.Dasovich R.Shapiro	0.1133	J.Dasovich J.Steffes	0.2440	T.Geaccone R.Haylett	0.0737
S.Beck S.White	0.0530	M.Taylor E.Sager	0.0218	J.Dasovich R.Sanders	0.1394	R.Haylett D.Fossum	0.0420

Beck = "Chief Operations Officer"

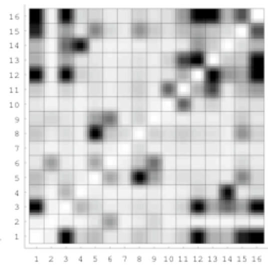
Dasovich = "Government Relations Executive"
Shapiro = "Vice Presidence of Regulatory Affairs"
Steffes = "Vice President of Government Affairs"

Discovering Role Similarity

Traditional SNA



ART



connection strength (A,B) =

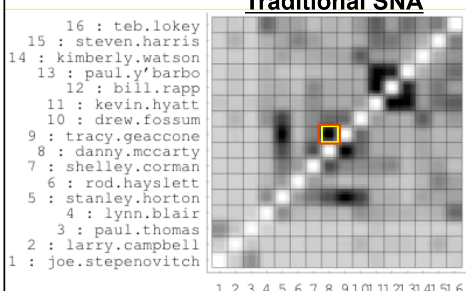
Similarity in recipients they sent email to

Similarity in authored topics, conditioned on recipient

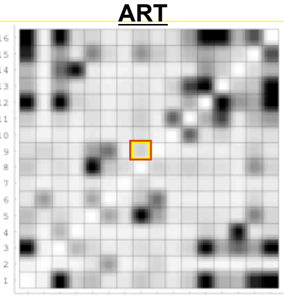
Discovering Role Similarity

Tracy Geaconne ⇔ Dan McCarty

Traditional SNA



ART



Similar (send email to same individuals)

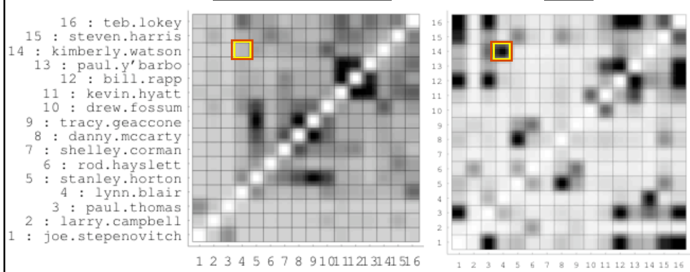
Different (discuss different topics)

Geaconne = "Secretary"
McCarty = "Vice President"

Discovering Role Similarity

Lynn Blair ⇔ Kimberly Watson

Traditional SNA **ART**



Different (send to different individuals)

Similar (discuss same topics)

Blair = "Gas pipeline logistics"
Watson = "Pipeline facilities planning"

What you should know

- Unsupervised dimension reduction using all features
  - Principle Components Analysis
    - Minimize reconstruction error
  - Singular Value Decomposition
    - Efficient PCA
  - Independent components analysis
  - Canonical correlation analysis
  - Probabilistic models with latent variables
- Supervised dimension reduction
  - Fisher Linear Discriminant
    - Project to n-1 dimensions to discriminate n classes
  - Hidden layers of Neural Networks
    - Most flexible, local minima issues
- LOTS of ways of combining discovery of latent features with classification tasks

Kernel Functions

- Kernel functions provide a way to manipulate data as though it were projected into a higher dimensional space, by operating on it in its original space
- This leads to efficient algorithms
- And is a key component of algorithms such as
  - Support Vector Machines
  - kernel PCA
  - kernel CCA
  - kernel regression
  - …

Linear Regression

Wish to learn f: X → Y, where X=<X_1, … X_n>, Y real-valued

Learn $\hat{f}(\mathbf{x}) = \sum_{i=1}^{N} x_i w_i = \underbrace{\langle \mathbf{x}, \mathbf{w} \rangle}_{\text{dot product}} = \mathbf{x}^{\mathbf{T}}\mathbf{w}$ transpose

bold → column vector or Matrix

$[x_1\, x_2 \cdots x_n]$ $\vec{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$

where $\mathbf{w} = \arg\min_{\mathbf{w}} \underbrace{\sum_{l=1}^{M} (y^l - {\mathbf{x}^{\mathbf{T}}}^l \mathbf{w})^2}_{l\text{th train example}} + \lambda \sum_{k}^{N} w_k^2$

$$\overbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix}}^{Y} - \overbrace{\begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_N^1 \\ & & & \\ x_1^M & x_2^M & \cdots & x_N^M \end{bmatrix}}^{X} \overbrace{\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}}^{w}$$

$$\|Y - Xw\|^2$$

Linear Regression

Wish to learn f: X → Y, where X=<X$_1$, … X$_n$>, Y real-valued

Learn $\hat{f}(\mathbf{x}) = \sum_{i=1}^{N} x_i w_i = \langle \mathbf{x}, \mathbf{w} \rangle = \mathbf{x}^{\mathbf{T}}\mathbf{w}$

where $\mathbf{w} = \arg\min_{\mathbf{w}} \sum_{l=1}^{M} (y^l - \mathbf{x}^{\mathbf{T}l}\mathbf{w})^2 + \lambda \sum_{k}^{N} w_k^2$

$$\mathbf{w} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \|\mathbf{w}\|^2$$

note l^{th} row of $\mathbf{X}$ is l^{th} training example $\mathbf{x}^{\mathbf{T}l}$

$$\|\mathbf{w}\|^2 = \sum_{k}^{N} w_k^2 = \|\mathbf{w}\|_2^2$$

Linear Regression

Wish to learn f: X → Y, where X=<X$_1$, … X$_n$>, Y real-valued

Learn $\hat{f}(\mathbf{x}) = \sum_{i=1}^{N} x_i w_i = \langle \mathbf{x}, \mathbf{w} \rangle = \mathbf{x}^{\mathbf{T}}\mathbf{w}$

where $\mathbf{w} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \|\mathbf{w}\|^2$

solve by taking derivative wrt w, setting to zero…

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\ \mathbf{X}^T\mathbf{y}$$

so: $\hat{f}(\mathbf{x}_{\mathbf{new}}) = \mathbf{x}_{\mathbf{new}}^{\mathbf{T}}\mathbf{w} = \mathbf{x}_{\mathbf{new}}^{\mathbf{T}}\ (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\ \mathbf{X}^T\mathbf{y}$