Stat 435 Intro to Statistical Machine Learning

Week 10: Time to recap! (and a little more SVM of course)

Richard Li

May 31, 2017

SVM: a summary

Binary response: $y_i \in \{-1, 1\}$

- Maximal margin classifier

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{maximize}}\, M \tag{9.9}$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1, \tag{9.10}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M \;\; \forall\, i = 1, \ldots, n. \tag{9.11}$$

- Support vector classifier
- Support vector classifier (alternative representation)
- Support vector machine

SVM: a summary

Binary response: $y_i \in \{-1, 1\}$

- Maximal margin classifier
- Support vector classifier

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M \tag{9.12}$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1, \tag{9.13}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i), \tag{9.14}$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C, \tag{9.15}$$

- Support vector classifier (alternative representation)
- Support vector machine

SVM: a summary

Binary response: $y_i \in \{-1, 1\}$

- Maximal margin classifier
- Support vector classifier

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M \tag{9.12}$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1, \tag{9.13}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i), \tag{9.14}$$

$$\epsilon_i \geq 0, \ \sum_{i=1}^{n} \epsilon_i \leq C, \tag{9.15}$$

- Support vector classifier (alternative representation)

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle, \tag{9.19}$$

- Support vector machine

SVM: a summary

Binary response: $y_i \in \{-1, 1\}$

- Maximal margin classifier
- Support vector classifier
- Support vector classifier (alternative representation)

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle x, x_i \rangle, \tag{9.19}$$

- Support vector machine

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K(x, x_i). \tag{9.23}$$

Relationship with logistic regression

- Yet another way to write support vector classifier

Relationship with logistic regression

- Yet another way to write support vector classifier

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \left\{ \sum_{i=1}^{n} \max\left[0, 1 - y_i f(x_i)\right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \qquad (9.25)$$

Relationship with logistic regression

- Yet another way to write support vector classifier

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \left\{ \sum_{i=1}^{n} \max\left[0, 1 - y_i f(x_i)\right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \qquad (9.25)$$

- The first part is a "loss function"
  - ...which is very similar to the loss function we minimize in logistic regression

Relationship with logistic regression

- Yet another way to write support vector classifier

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \left\{ \sum_{i=1}^{n} \max\left[0, 1 - y_i f(x_i)\right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \qquad (9.25)$$

- The first part is a “loss function”
  - ...which is very similar to the loss function we minimize in logistic regression
- The second part is a ridge penalty!

Overfitting, curse of dimensionality

- In practice, SVM usually used when data is high dimensional
- It tends to be resistant to overfitting because of the regularization

Overfitting, curse of dimensionality

- In practice, SVM usually used when data is high dimensional
- It tends to be resistant to overfitting because of the regularization
- BUT, successful implementation of SVM still depends on careful tunning of

Overfitting, curse of dimensionality

- In practice, SVM usually used when data is high dimensional
- It tends to be resistant to overfitting because of the regularization
- BUT, successful implementation of SVM still depends on careful tunning of
  - C

Overfitting, curse of dimensionality

- In practice, SVM usually used when data is high dimensional
- It tends to be resistant to overfitting because of the regularization
- BUT, successful implementation of SVM still depends on careful tunning of
  - C
  - kernel (imagine you want to cut the skins of an orange by a linear plane)

The kernel trick

- Yes that is the formal name...

The kernel trick

- Yes that is the formal name...
- A general approach to turn linear functions into nonlinear functions

The kernel trick

- Yes that is the formal name...
- A general approach to turn linear functions into nonlinear functions
- What it does equivalently: transform data into higher dimensional space (without explicitly calculating/storing the transformation)

The kernel trick

- Yes that is the formal name...
- A general approach to turn linear functions into nonlinear functions
- What it does equivalently: transform data into higher dimensional space (without explicitly calculating/storing the transformation)
    - Instead of $\langle x_i, x_j \rangle$, $K(x_i, x_j)$ can be thought of as $\langle x_i', x_j' \rangle$ where x' is of a higher dimension than x.

The kernel trick

- Yes that is the formal name...
- A general approach to turn linear functions into nonlinear functions
- What it does equivalently: transform data into higher dimensional space (without explicitly calculating/storing the transformation)
  - Instead of $\langle x_i, x_j \rangle$, $K(x_i, x_j)$ can be thought of as $\langle x_i', x_j' \rangle$ where x' is of a higher dimension than x.
  - e.g., a polynomial kernel

    $K([a, b], [c, d]) = (ac + bd + 5)^2 = a^2c^2 + 2acbd + 10ac + 10cd + b^2d^2 + 25$

The kernel trick

- Yes that is the formal name...
- A general approach to turn linear functions into nonlinear functions
- What it does equivalently: transform data into higher dimensional space (without explicitly calculating/storing the transformation)
  - Instead of $\langle x_i, x_j \rangle$, $K(x_i, x_j)$ can be thought of as $\langle x_i', x_j' \rangle$ where x' is of a higher dimension than x.
  - e.g., a polynomial kernel

    $$K([a, b], [c, d]) = (ac+bd+5)^2 = a^2c^2+2acbd+10ac+10cd+b^2d^2+25$$

  - equivalently,

$$K([a, b], [c, d]) = \langle \begin{bmatrix} a^2 \\ b^2 \\ \sqrt{2}ab \\ \sqrt{10}a \\ \sqrt{10}b \\ 5 \end{bmatrix}, \begin{bmatrix} c^2 \\ d^2 \\ \sqrt{2}cd \\ \sqrt{10}c \\ \sqrt{10}d \\ 5 \end{bmatrix} \rangle$$

The kernel trick

- Yes that is the formal name...
- A general approach to turn linear functions into nonlinear functions
- What it does equivalently: transform data into higher dimensional space (without explicitly calculating/storing the transformation)
  - Instead of $\langle x_i, x_j \rangle$, $K(x_i, x_j)$ can be thought of as $\langle x'_i, x'_j \rangle$ where x' is of a higher dimension than x.
  - e.g., a polynomial kernel

    $K([a, b], [c, d]) = (ac+bd+5)^2 = a^2c^2+2acbd+10ac+10cd+b^2d^2+25$

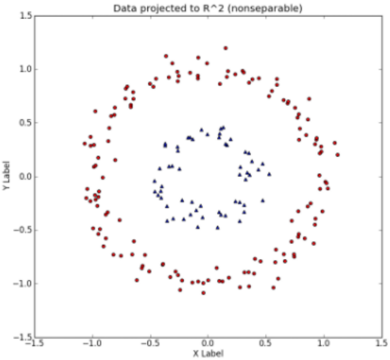
  - equivalently,

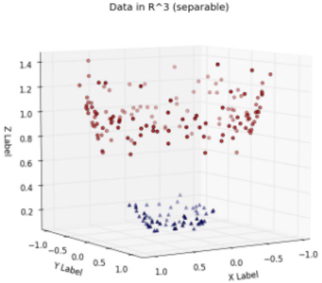
$$K([a, b], [c, d]) = \langle \begin{bmatrix} a^2 \\ b^2 \\ \sqrt{2}ab \\ \sqrt{10}a \\ \sqrt{10}b \\ 5 \end{bmatrix}, \begin{bmatrix} c^2 \\ d^2 \\ \sqrt{2}cd \\ \sqrt{10}c \\ \sqrt{10}d \\ 5 \end{bmatrix} \rangle$$

which transforms our data in $\mathbb{R}^2$ to $\mathbb{R}^5$

The kernel trick

$[x_1, x_2] \rightarrow [x_1, x_2, x_1^2 + x_2^2]$





But You don't need to look for such transformations directly (difficult in high dimensions!), thus "trick" as in "the kernel trick"

The kernel trick

- It can be used in regression
- It can be used in PCA
- It can be used in KNN
- In many algorithms where you see $x^T x$, chances are the kernel trick can be applied...

One of the last general methods in this course!

So now let's look back at what we have learned in this class

So now let's look back at what we have learned in this class

What you have learned

First of all,

- Believe it or not, you almost finished a book with 400+ pages!

What you have learned

First of all,

- Believe it or not, you almost finished a book with 400+ pages!

Now it's a good time to review how we get here...

- What is statistical learning?

What you have learned

First of all,

- Believe it or not, you almost finished a book with 400+ pages!

Now it's a good time to review how we get here...

- What is statistical learning?
    - Tools for *understanding data*

What you have learned

First of all,

- Believe it or not, you almost finished a book with 400+ pages!

Now it's a good time to review how we get here...

- What is statistical learning?
    - Tools for *understanding data*
- In what kind of problems do we need to understand data?

What you have learned

First of all,

- Believe it or not, you almost finished a book with 400+ pages!

Now it's a good time to review how we get here...

- What is statistical learning?
  - Tools for *understanding data*
- In what kind of problems do we need to understand data?
  - Predictive, an outcome we want to predict
  - Descriptive, understand relationship and structures in data

What you have learned

First of all,

- Believe it or not, you almost finished a book with 400+ pages!

Now it's a good time to review how we get here...

- What is statistical learning?
  - Tools for *understanding data*
- In what kind of problems do we need to understand data?
  - Predictive, an outcome we want to predict
  - Descriptive, understand relationship and structures in data
- What tools have we learned?

Tools

Is their an outcome?

Tools

Is their an outcome?

- Supervised:
- Unsupervised:

Tools

Is their an outcome?

- Supervised:
- Unsupervised:

Is the outcome qualitative or quantitative?

Tools

Is their an outcome?

- Supervised:
- Unsupervised:

Is the outcome qualitative or quantitative?

- Linear regression, more regression (stepwise, ridge, lasso, PCR, PLS, etc.), splines, regression trees, GAM, ...
- KNN, LDA, QDA, logistic regression, SVM, decision trees, ...

Tools

Is their an outcome?

- Supervised:
- Unsupervised:

Is the outcome qualitative or quantitative?

- Linear regression, more regression (stepwise, ridge, lasso, PCR, PLS, etc.), splines, regression trees, GAM, ...
- KNN, LDA, QDA, logistic regression, SVM, decision trees, ...

Do you want to reduce or enrich the complexity?

- Regularization: stepwise, ridge, and lasso regression, ...
- Non-linear: polynomial regression, splines, local regression, GAM, ...

Tools

Is their an outcome?

- Supervised:
- Unsupervised:

Is the outcome qualitative or quantitative?

- Linear regression, more regression (stepwise, ridge, lasso, PCR, PLS, etc.), splines, regression trees, GAM, ...
- KNN, LDA, QDA, logistic regression, SVM, decision trees, ...

Do you want to reduce or enrich the complexity?

- Regularization: stepwise, ridge, and lasso regression, ...
- Non-linear: polynomial regression, splines, local regression, GAM, ...

Do we care about model interpretability or just prediction accuracy

- See figure 2.7 of text book (page 25)

Common considerations

THE Bias-Variance Trade-Off

- Model selection: Cross-validation
- Model selection: AIC, BIC, C_p, adjusted R^2, ...

Common considerations

THE Bias-Variance Trade-Off

- Model selection: Cross-validation
- Model selection: AIC, BIC, C_p, adjusted R^2, ...
- Reducing variance: bagging, boosting, random forest

Common considerations

THE Bias-Variance Trade-Off

- Model selection: Cross-validation
- Model selection: AIC, BIC, C_p, adjusted R^2, ...
- Reducing variance: bagging, boosting, random forest

Related reoccurring topics (in high dimensional data)

- Simpler model is more favorable: *Penalization*
- More flexible model is more favorable (if can be computed efficiently): *Kernel methods*

A quick ride in statistical learning land

But incomplete and not meant as the full material for final!

Bias-Variance Trade-off

Our journey begins with a broad goal

Bias-Variance Trade-off

Our journey begins with a broad goal

- MSE
- Can you still derive it? (*midterm problem*)

Bias-Variance Trade-off

Our journey begins with a broad goal

- MSE
- Can you still derive it? (*midterm problem*)
- Where does variance and bias come from? (*Week5*)

Bias-Variance Trade-off

Our journey begins with a broad goal

- MSE
- Can you still derive it? (*midterm problem*)
- Where does variance and bias come from? (*Week5*)
- Implications, e.g., correct model does not always give best MSE (*midterm, Week5*)

Bias-Variance Trade-off

Our journey begins with a broad goal

- MSE
- Can you still derive it? (*midterm problem*)
- Where does variance and bias come from? (*Week5*)
- Implications, e.g., correct model does not always give best MSE (*midterm, Week5*)
- Bayes classifier / Bayes error rate (*HW1, Week1*)

Linear regression

Then there came a ‘simple’ problem

Linear regression

Then there came a 'simple' problem

- Derivation (*HW2*)

Linear regression

Then there came a 'simple' problem

- Derivation (*HW2*)
- Interpretation, tests, diagnostics (*week2*)

Linear regression

Then there came a 'simple' problem

- Derivation (*HW2*)
- Interpretation, tests, diagnostics (*week2*)
- Extensions (*week2*)
  - Interaction (interpretation!)
  - Qualitative predictors

Classification

Then a little more complication with the outcome

Classification

Then a little more complication with the outcome

- Logistic regression
    - Derivation

Classification

Then a little more complication with the outcome

- Logistic regression
  - Derivation
- LDA and QDA (*HW3*)
  - Decision boundary (*week5*)

Classification

Then a little more complication with the outcome

- Logistic regression
  - Derivation
- LDA and QDA (*HW3*)
  - Decision boundary (*week5*)
  - Extensions (*Midterm*)

Classification

Then a little more complication with the outcome

- Logistic regression
  - Derivation
- LDA and QDA (*HW3*)
  - Decision boundary (*week5*)
  - Extensions (*Midterm*)
  - Bayes theorem / Bayes error rate (*HW1, Week3, Midterm*)

Classification

Then a little more complication with the outcome

- Logistic regression
  - Derivation
- LDA and QDA (*HW3*)
  - Decision boundary (*week5*)
  - Extensions (*Midterm*)
  - Bayes theorem / Bayes error rate (*HW1, Week3, Midterm*)
- KNN
  - Intuition

Classification

Then a little more complication with the outcome

- Logistic regression
    - Derivation
- LDA and QDA (*HW3*)
    - Decision boundary (*week5*)
    - Extensions (*Midterm*)
    - Bayes theorem / Bayes error rate (*HW1, Week3, Midterm*)
- KNN
    - Intuition
    - Why it might not be a good idea sometimes? (*HW6*)

Classification

Then a little more complication with the outcome

- Logistic regression
    - Derivation
- LDA and QDA (*HW3*)
    - Decision boundary (*week5*)
    - Extensions (*Midterm*)
    - Bayes theorem / Bayes error rate (*HW1, Week3, Midterm*)
- KNN
    - Intuition
    - Why it might not be a good idea sometimes? (*HW6*)

Of course, later we learned something more complicated

Classification

Then a little more complication with the outcome

- Logistic regression
  - Derivation
- LDA and QDA (*HW3*)
  - Decision boundary (*week5*)
  - Extensions (*Midterm*)
  - Bayes theorem / Bayes error rate (*HW1, Week3, Midterm*)
- KNN
  - Intuition
  - Why it might not be a good idea sometimes? (*HW6*)

Of course, later we learned something more complicated

- (GAM)
- Classification tree
- SVM (related to logistic regression! Sec 9.5 textbook)

Resampling

Then we did a brief detour to learn about some magic tricks of creating testing data from nowhere (and had a midterm, eww!)

Resampling

Then we did a brief detour to learn about some magic tricks of creating testing data from nowhere (and had a midterm, eww!)

- Cross validation

Resampling

Then we did a brief detour to learn about some magic tricks of creating testing data from nowhere (and had a midterm, eww!)

- Cross validation
  - Different approaches (*week4*)

Resampling

Then we did a brief detour to learn about some magic tricks of creating testing data from nowhere (and had a midterm, eww!)

- Cross validation
  - Different approaches (*week4*)
  - How they compare? (*week6*)

Resampling

Then we did a brief detour to learn about some magic tricks of creating testing data from nowhere (and had a midterm, eww!)

- Cross validation
  - Different approaches (*week4*)
  - How they compare? (*week6*)
- Bootstrap

Resampling

Then we did a brief detour to learn about some magic tricks of creating testing data from nowhere (and had a midterm, eww!)

- Cross validation
  - Different approaches (*week4*)
  - How they compare? (*week6*)
- Bootstrap
  - Used for getting variance of estimator

Resampling

Then we did a brief detour to learn about some magic tricks of creating testing data from nowhere (and had a midterm, eww!)

- Cross validation
  - Different approaches (*week4*)
  - How they compare? (*week6*)
- Bootstrap
  - Used for getting variance of estimator
  - Used in bagging

Resampling

Then we did a brief detour to learn about some magic tricks of creating testing data from nowhere (and had a midterm, eww!)

- Cross validation
  - Different approaches (*week4*)
  - How they compare? (*week6*)
- Bootstrap
  - Used for getting variance of estimator
  - Used in bagging
  - Properties (*HW8*)

Model selection/regularization

Then we encountered the dragon of high-dimensionality

Model selection/regularization

Then we encountered the dragon of high-dimensionality

- Subset selection (*HW4*)
  - Algorithms, how do MSE change at each step

Model selection/regularization

Then we encountered the dragon of high-dimensionality

- Subset selection (*HW4*)
    - Algorithms, how do MSE change at each step
- Shrinkage (ridge and lasso) (*week6, week7, week8, HW4, HW5*)
    - Derivation

Model selection/regularization

Then we encountered the dragon of high-dimensionality

- Subset selection (*HW4*)
    - Algorithms, how do MSE change at each step
- Shrinkage (ridge and lasso) (*week6, week7, week8, HW4, HW5*)
    - Derivation
    - Properties (exact zero, degrees of freedom, ...)

Model selection/regularization

Then we encountered the dragon of high-dimensionality

- Subset selection (*HW4*)
  - Algorithms, how do MSE change at each step
- Shrinkage (ridge and lasso) (*week6, week7, week8, HW4, HW5*)
  - Derivation
  - Properties (exact zero, degrees of freedom, ...)
- Dimension reduction (*week8*)
  - PCR and PLS
  - Intuitive understanding

Model selection/regularization

Then we encountered the dragon of high-dimensionality

- Subset selection (*HW4*)
  - Algorithms, how do MSE change at each step
- Shrinkage (ridge and lasso) (*week6, week7, week8, HW4, HW5*)
  - Derivation
  - Properties (exact zero, degrees of freedom, ...)
- Dimension reduction (*week8*)
  - PCR and PLS
  - Intuitive understanding
- Why high-dimension is difficult?
  - curse of dimensionality (*week6, HW6*)

Model selection/regularization

Then we encountered the dragon of high-dimensionality

- Subset selection (*HW4*)
    - Algorithms, how do MSE change at each step
- Shrinkage (ridge and lasso) (*week6, week7, week8, HW4, HW5*)
    - Derivation
    - Properties (exact zero, degrees of freedom, ...)
- Dimension reduction (*week8*)
    - PCR and PLS
    - Intuitive understanding
- Why high-dimension is difficult?
    - curse of dimensionality (*week6, HW6*)

Nonlinear additive models

Then we took a step back and started thinking about non-linearity

Nonlinear additive models

Then we took a step back and started thinking about non-linearity

- Step functions, polynomial regressions, and regression splines (*HW6*)

Nonlinear additive models

Then we took a step back and started thinking about non-linearity

- Step functions, polynomial regressions, and regression splines (*HW6*)
- Smoothing splines (*HW6*)

Nonlinear additive models

Then we took a step back and started thinking about non-linearity

- Step functions, polynomial regressions, and regression splines (*HW6*)
- Smoothing splines (*HW6*)
- Generalized Additive Models (*HW7*)

Trees

Then we learned a completely new set of exotic skills that seem so simple but so difficult at the same time

Trees

Then we learned a completely new set of exotic skills that seem so simple but so difficult at the same time

- Regression and decision trees (*HW7, week9*)

Trees

Then we learned a completely new set of exotic skills that seem so simple but so difficult at the same time

- Regression and decision trees (*HW7, week9*)
- Bagging, boosting, and random forest(*week9*)

SVM and unsupervised learning

Then there's the recent stuff, and they should still be fresh in your memory

There's a lot lot more to statistical learning

Three-Eyed Raven: *You won't be here forever. You won't be an old man in a tree. But before you leave, you must learn.*

Bran Stark: *Learn what?*

Three-Eyed Raven: *Hmm let's see, bias-variance tradeoff, regression, classification, cross-validation, regularization, model selection, dimension reduction splines, GAMs, trees, bagging, boosting, random forests, support vector machines, PCA, LDA, QDA, ...*

(Again, true script from Game of Thrones: Oathbreaker (#6.3))

There's a lot lot more to statistical learning

Three-Eyed Raven: *You won't be here forever. You won't be an old man in a tree. But before you leave, you must learn.*

Bran Stark: *Learn what?*

Three-Eyed Raven: *Hmm let's see, bias-variance tradeoff, regression, classification, cross-validation, regularization, model selection, dimension reduction splines, GAMs, trees, bagging, boosting, random forests, support vector machines, PCA, LDA, QDA, ...*

(Again, true script from Game of Thrones: Oathbreaker (#6.3))

Course evaluation closing on Friday!

https://uw.iasystem.org/survey/174515

Thank you!