Support Vector Machines

Michael Rabadi

New York University

michael.rabadi@nyu.edu

May 19, 2015

Outline

1. Poll
2. Introduction to Machine Learning
   - Overview
   - Goal of Machine Learning (supervised)
3. Notion of Generalization
   - Complexity
   - ERM and SRM
   - Generalization Bounds
4. Cross Validation
5. Support Vector Machines - Geometric Intuitions
6. Mathematical Formalism - Separable Case
7. SVM Tutorial
8. Support Vector Machines - Geometric Intuitions
9. Appendix

Decide on structure

- Who wants to present?
- Google Doc - please put name down
- Topic interest

What is Machine Learning?

- Methods for using data to generate statistical models.
- Data-driven
- Study and design of algorithms
- ML Theory: complexity analysis; learning guarantees

Areas of ML

- Classification: assign a label to examples (object recognition)
- Regression: predict value of item
- Clustering: split up data to extract "structure"
- Dimensionality Reduction: find low-dimensional manifold within a high-dimensional space.

Definitions and Jargon

- Example: data point
- Features: attributes of a data point (pixels in image, for example)
- Labels: Category or value associated to data point
- Training data: Labeled or unlabeled - use this to train your model
- Validation data: Labeled - use this to adjust hyper-parameters
- Test data: Labeled, but not seen - only use this to test your model!

Definitions and Jargon

- Spaces: where you draw samples from, typically an input space $\mathcal{X}$ and output space $\mathcal{Y}$
- Loss function: $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
  - Cost of predicting $\hat{y}$ instead of y
- Hypothesis set: $H \subseteq \mathcal{Y}^{\mathcal{X}}$
  - Subset of functions from which a hypothesis is selected

Definitions: types of errors

- True error (AKA Baye's error): the error that is inherent to the system - you can never get rid of this! Related to the inherent noise of system.

$$R^* = \inf_h R(h)$$

  - Note: $\mathrm{E}[\mathrm{noise}(x)] = R^*$

- Empirical error: the error that you measure for the model h that you have learned, given your sample S.

$$\hat{R}(h) = \frac{1}{m}\sum_{i=1}^{m} L(h(x_i), y_i)$$

- Generalization error: The error of your hypothesis, in general. How well your model will predict data that is has never seen.

$$R(h) = \mathrm{E}_{(x,y)\sim D}[L(h(x), y)]$$

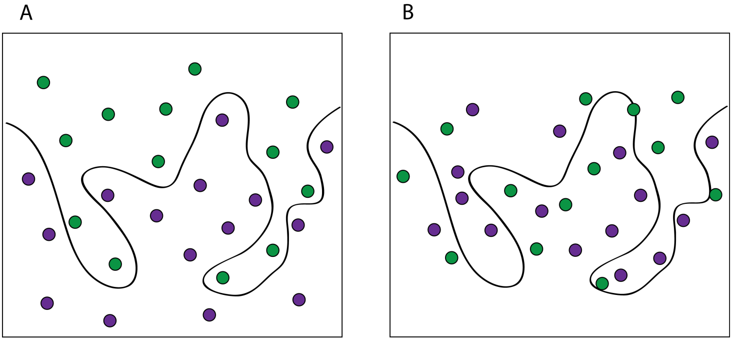
Problem

- Given sample S of size m:

$$S = ((x_1, y_1), ..., (x_m, y_m)).$$

- Goal: find a hypothesis $h \in H$ with small *generalization error*
- **WE ONLY CARE ABOUT GENERALIZATION ERROR**

Overfitting



- Figure A - training set performance
- Figure B - test set performance
- $\hat{R}(h) = 0$ on training set
- 16 mistakes on test set

Outline

1. Poll
2. Introduction to Machine Learning
   - Overview
   - Goal of Machine Learning (supervised)
3. Notion of Generalization
   - Complexity
   - ERM and SRM
   - Generalization Bounds
4. Cross Validation
5. Support Vector Machines - Geometric Intuitions
6. Mathematical Formalism - Separable Case
7. SVM Tutorial
8. Support Vector Machines - Geometric Intuitions
9. Appendix

Complexity

- Multiple definitions, but should relate to how much a hypothesis can (over)fit data.
- Complexity describes a hypothesis h (i.e., your model)
- We want a model that is complex enough to describe the data, but not so complex that we overfit - exact notion of Occam's Razor
- Notion of model complexity is ubiquitous - AIC, BIC, corss-validation

Empirical Risk Minimization

- Select a hypothesis set (assume a set of models)
- Choose $h \in H$ such that:

$$h = \mathrm{argmin}_{h \in H} \hat{R}(h)$$

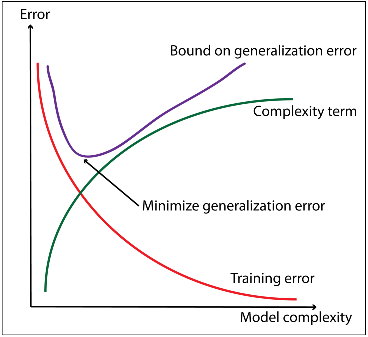
- Commonly used - but doesn't tell you anything about $R(h)$.
- No theoretical justification

Structural Risk Minimization

- Select a hypothesis such that:

$$h = \operatorname{argmin}_{h \in H} \hat{R}(h) + \text{complexity}(H, m)$$

- Theoretical justification - take model complexity into account



Generalization Bounds

- Upper bound on error of model

$$\Pr[|R(h) - \hat{R}(h)| > \epsilon] \leq \text{bound}$$

- Goal is to select model that has smallest generalization bound
- Generalization bound depends on model complexity (i.e., Rademacher complexity or VC dimension).
- Natural and theoretically justified
- Detailed description of generalization bounds and how to derive them are outside the scope of this seminar.... But we will discuss another time!

Outline

1. Poll
2. Introduction to Machine Learning
   - Overview
   - Goal of Machine Learning (supervised)
3. Notion of Generalization
   - Complexity
   - ERM and SRM
   - Generalization Bounds
4. Cross Validation
5. Support Vector Machines - Geometric Intuitions
6. Mathematical Formalism - Separable Case
7. SVM Tutorial
8. Support Vector Machines - Geometric Intuitions
9. Appendix

Cross Validation

- Shuffle data
- Leave 10% of data on the side - this will be the test set.
- Split up the remaining data into 10 equal-size disjoint sets.
    - Combine 9 for training set. Last one is the validation set.
- Train model on the training set. Measure performance on validation set.
- Switch out a training set with the validation set and repeat until each of the 10 sets have been used for validation
- Repeat for different parameters
- Choose model (with parameters) that was best on validation - confirm with the test set

Never Cross Validate with the Test Set

NEVER CROSS VALIDATE WITH THE TEST SET!!!

- Cross validating with the test set is *cheating*
- This is commonly done, usually by mistake
- Common, subtle example: "We left 10% of our data aside. We trained our model with the rest of the data and cross validated with the test set. This is the performance of our model on the test data...."

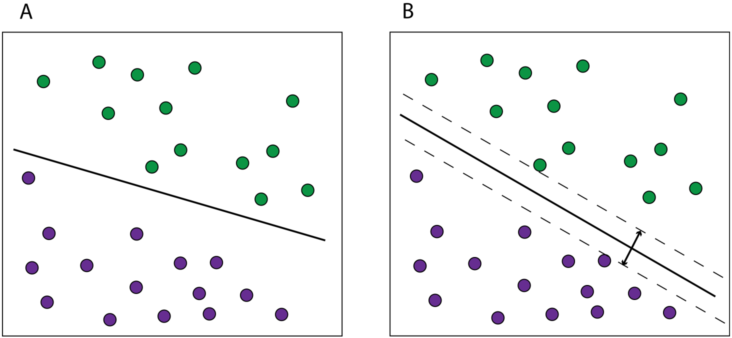
How to select a good model

- AIC and BIC - penalty for ”model complexity”
  - AIC only for parametric models and proof depends on assumption that the true model is within the set that you are comparing. [Cavanaugh, 2012]
  - BIC also assumes that the best model is in comparison set [Dudley]
  - In practice, assumption might not matter - but no theoretical justification.
- Cross Validation - use data to test model complexity:
  - Poor CV performance means model is either too complex or not complex enough
  - Good CV performance implied good performance
  - Proof that the error of a model returned via cross validation will be close to value of a model obtained through SRM [proof in appendix]

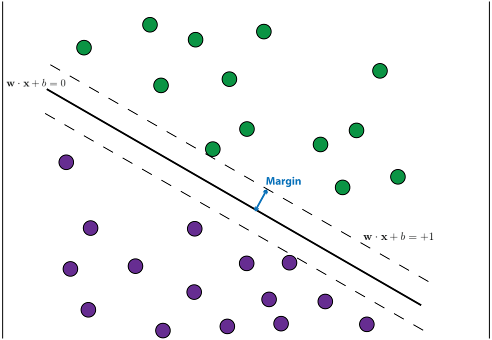
Outline

1. Poll
2. Introduction to Machine Learning
   - Overview
   - Goal of Machine Learning (supervised)
3. Notion of Generalization
   - Complexity
   - ERM and SRM
   - Generalization Bounds
4. Cross Validation
5. Support Vector Machines - Geometric Intuitions
6. Mathematical Formalism - Separable Case
7. SVM Tutorial
8. Support Vector Machines - Geometric Intuitions
9. Appendix

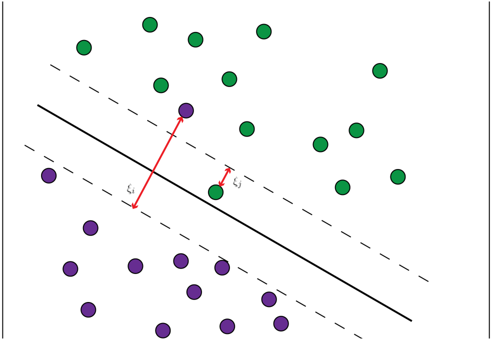
Support Vector Machines - Separable Case



Support Vector Machines - Separable Case


$\mathbf{w} \cdot \mathbf{x} + b = 0$
Margin
$\mathbf{w} \cdot \mathbf{x} + b = +1$

Support Vector Machines - Non-separable Case



Support Vector Machines - Binary Classification

- Samples drawn i.i.d. according to an unknown distribution D
- $S = ((x_1, y_1), ..., (x_m, y_m)) \in X \times \{-1, +1\}$
- Find model $h \in H$ such that $h : X \mapsto \{-1, +1\}$

Marginal Hyperplanes

Equation for hyperplane in $\mathbb{R}^N$:

$$\mathbf{x} \cdot \mathbf{w} + b = 0$$

where $\mathbf{w} \in \mathbb{R}^N$ is a vector normal to the hyperplane and b is a scalar. Thus, for any hyperplane that does not pass through any samples:

$$\min_{(\mathbf{x},y)\in S} |\mathbf{w} \cdot \mathbf{x} + b| = 1$$

Thus, the hyperplane correctly classifies a training point $\mathbf{x}_i$ when $\mathbf{w} \cdot \mathbf{x}_i + b$ has the same sign as y_i, in other words, when $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$.

Margin

Let ρ denote the margin:

$$\rho = \min_{(\mathbf{x},y)\in S} \frac{|\mathbf{w}\cdot\mathbf{x}+b|}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$$

Notice that maximizing the margin ρ is equivalent to minimizing $\|\mathbf{w}\|$.

Optimal Hyperplane: Separable Case

We will want to find the maximum margin, subject to the following condition:

$$\rho = \max_{\mathbf{w},b:y_i(\mathbf{w}\cdot\mathbf{x}+b)\geq 0)} \min_{i\in[1,m]} \frac{|\mathbf{w}\cdot\mathbf{x}_i + b|}{\|\mathbf{w}\|}$$

Which as we've seen from above is equivalent to:

$$\rho = \max_{\mathbf{w},b:y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1} \frac{1}{\|\mathbf{w}\|}$$

Optimization: Separable Case

The optimization problem can be rewritten as follows:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

subject to: $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \forall i \in [1, m]$

- Admits an infinitely differentiable, strictly convex objective function:
- $F : \mathbf{w} \mapsto \frac{1}{2}\|\mathbf{w}|^2$
- $\nabla_{\mathbf{w}}(f) = \mathbf{w}$ and $\nabla^2 f(\mathbf{w}) = \mathbf{I}$
- Just throw into your favorite quadratic programming optimizer and find the global minimum.

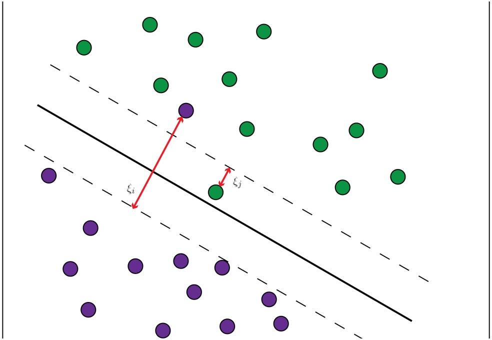
Support Vectors

- The constraints can be rewritten in the Lagrangian form.
- Let $\alpha_i \geq 0, i \in [1, m]$ be the Lagrange variables for the optimization criteria

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1]$$

- $\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$: $\mathbf{w}$ is linear combination of training set
- $\sum_{i=1}^{m} \alpha_i y_i = 0$
- $\alpha_i = 0 \vee y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1$: support vectors lie on marginal hyperplane
- Support vectors define the optimal hyperplane - hence SVM
- Solution only depends on support vectors!

Support Vector Machines - Non-separable Case



SVM - non-separable case

- Usually, data cannot be linearly separated
- Introduce a so called slack variable ξ_i for optimization
- Optimization constraint relaxed:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i$$

- Vector $\mathbf{x}_i$ with $\xi_i > 0$ is an *outlier*
- But too much slack and we can't find a hyperplane
- Conflicting objectives: large margin (more outliers) vs fewer outliers (small margin)

Optimization: non-separable case

Let $C \geq 0$ be a parameter chosen via cross-validation. C determines the trade-off between the maximum margin size and the slack penalty:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$

subject to $y_i(\mathbf{w}\cdot\mathbf{x}+b) \geq 1-\xi_i \wedge \xi_i \geq 0, i \in [1,m]$

Generalization guarantee Mohri, 2012

The empirical margin loss tells us how many points are mislabeled given some margin ρ is defined as follows (true definition is finer, but is for another time):

$$\hat{R}_\rho \leq \frac{1}{m} \sum_{i=1}^{m} 1_{y_i h(x_i) \leq \rho}$$

Let H denote a set of functions from which h, our model, is drawn. Fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following hold for all $h \in H$:

$$R(h) \leq \hat{R}_\rho(h) + \frac{2}{\rho} \mathfrak{R}_m(H) + \sqrt{\frac{log \frac{1}{\delta}}{2m}}$$

where $\mathfrak{R}_m(H)$ denotes the Rademacher complexity for the hypothesis set over the sample. The Rademacher complexity measures how well a hypothesis can fit random noise. We can talk about it more in another seminar.

Generalization guarantee notes

$$R(h) \leq \hat{R}_\rho(h) + \frac{2}{\rho}\mathfrak{R}_m(H) + \sqrt{\frac{log\,\frac{1}{\delta}}{2m}}$$

- Larger margin decreases complexity term
- But larger margin leads to higher error!
- Use cross validation to find optimal tradeoff

Outline

1. Poll
2. Introduction to Machine Learning
   - Overview
   - Goal of Machine Learning (supervised)
3. Notion of Generalization
   - Complexity
   - ERM and SRM
   - Generalization Bounds
4. Cross Validation
5. Support Vector Machines - Geometric Intuitions
6. Mathematical Formalism - Separable Case
7. SVM Tutorial
8. Support Vector Machines - Geometric Intuitions
9. Appendix

MATLAB Tutorial

Appendix

Cross validation

Theorem: Let $(H_k)_{k\in\mathbb{N}}$ be a countable sequence of hypothesis sets with increasing complexities. Assume that the cross-validation (CV) solution is obtained as follows. A learner receives an i.i.d. sample S of size $m \geq 1$. It randomly divides S into a sample S_1 of size $(1-\alpha)m$ and sample S_2 of size αm, where α is in $(0, 1)$ and is small. S_1 is used for training, S_2 for validation. For any $k \in \mathbb{N}$, let $\hat{h}_k$ denote the ERM run on S_1 using hypothesis set H_k. The learner then uses sample S_2 to return the CV solution $f_{\mathrm{CV}} = \mathrm{argmin}_k \in N\hat{R}_{S_2}(\hat{h}_k)$. Also let $R(f_{\mathrm{SRM}}, S_1)$ be the generalization error of the SRM solution using a sample S_1 of size $(1-\alpha m)$ and $R(f_{\mathrm{CV}}, S)$ the generalization error of the cross-validation solution using a sample S of size m. Then, for any $\delta > 0$, with probability at least $1-\delta$ the following holds:

$$R(f_{\mathrm{CV}}, S) - R(f_{\mathrm{SRM}}, S_1) \leq 2\sqrt{\frac{\log\frac{4}{\delta}}{2\alpha m}} + 2\sqrt{\frac{\log\max(k(f_{\mathrm{CV}}), k(f_{\mathrm{SRM}}))}{\alpha m}},$$

where, for any h, $k(h)$ denotes the smallest index of a hypothesis set containing h.

Cross validation proof

Proof. By the union bound, we have

$$\begin{aligned}
&\Pr\left(\sup_{k\geq 1}|R(\hat{h}_k)-\hat{R}_{S_2}(\hat{h}_k)|>\epsilon+\sqrt{\frac{\log k}{\alpha m}}\right)\\
&\leq\sum_{k=1}^{\infty}\Pr\left(|R(\hat{h}_k)-\hat{R}_{S_2}(\hat{h}_k)|>\epsilon+\sqrt{\frac{\log k}{\alpha m}}\right)\\
&=\sum\mathrm{E}\Big[\Pr\Big(|R(\hat{h}_k)-\hat{R}_{S_2}(\hat{h}_k)|>\epsilon+\sqrt{\frac{\log k}{\alpha m}}|S_1\Big)\Big].
\end{aligned}$$

Since $\hat{h}_k$ is conditioned on S_1 and sample S_2 is independent from sample S_1, the following holds by Hoeffding's inequality:

$$\begin{aligned}
&\Pr\left(|R(\hat{h}_k)-\hat{R}_{S_2}(\hat{h}_k)|>\epsilon+\sqrt{\frac{\log k}{\alpha m}}|S_1\right)\\
&\leq 2\exp\left(-2\alpha m\Big(\epsilon+\sqrt{\frac{\log k}{\alpha m}}\Big)^2\right)\leq\exp(-2\alpha m\epsilon^2-2\log k)\\
&=\frac{1}{k^2}\exp(-2\alpha m\epsilon^2)
\end{aligned}$$

Cross validation proof cont.

Thus:

$$\Pr\left(\sup_{k\geq 1}|R(\hat{h}_k) - \hat{R}_{S_2}(\hat{h}_k)| > \epsilon + \sqrt{\frac{\log k}{\alpha m}}\right) \leq 4\exp(2\alpha m \epsilon^2).$$

Given this bound, then with probability at least $1-\delta$:

$$\begin{aligned}
R(f_{\mathrm{CV}}, S) &\leq \hat{R}_{S_2}(f_{\mathrm{CV}}) + \sqrt{\frac{\log\frac{4}{\delta}}{2\alpha m}} + \sqrt{\frac{\log(k(f_{\mathrm{CV}}))}{\alpha m}}\\
&\leq \hat{R}_{S_2}(f_{\mathrm{SRM}}) + \sqrt{\frac{\log\frac{4}{\delta}}{2\alpha m}} + \sqrt{\frac{\log(k(f_{\mathrm{CV}}))}{\alpha m}}\\
&\leq R(f_{\mathrm{SRM}}, S_1) + 2\sqrt{\frac{\log\frac{4}{\delta}}{2\alpha m}} + \sqrt{\frac{\log(k(f_{\mathrm{CV}}))}{\alpha m}} + \sqrt{\frac{\log(k(f_{\mathrm{SRM}}))}{\alpha m}}\\
& R(f_{\mathrm{SRM}}, S_1) + 2\sqrt{\frac{\log\frac{4}{\delta}}{2\alpha m}} + 2\sqrt{\frac{\log(\max(k(f_{\mathrm{CV}}), k(f_{\mathrm{SRM}}))}{\alpha m}} \quad q.e.d.
\end{aligned}$$

Cross validation notes

- The cross validation solution will be close to the SRM solution!
- Training on $(1 - \alpha)m$ points could be poor in some bad cases

AIC and BIC

Let k be the number of parameters in a model and n be the sample size. Let L be the max value of the likelihood function such that $L = \Pr(x|\mathrm{Parameters}, \mathrm{Model})$.

- $\mathrm{AIC} = 2k - 2\ln(L)$
- $\mathrm{BIC} = k \cdot \ln(n) - 2 \cdot \ln(L)$

Resources

- Cavanaugh, J. (2012). The Akaike Information Criterion.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.
- Dudley, R. M. (2014). The Bayes Information Criterion (BIC).
- Koltchinskii, V., & Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. Annals of Statistics, 1-50.
- Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of machine learning. MIT press.
- Vapnik, V. N., & Vapnik, V. (1998). Statistical learning theory (Vol. 1). New York: Wiley.