Recitation 6

Review for Midterm

Vishakh

CDS

March 2, 2022

Announcement

- Midterm next week
- HW 3 is due Friday
- Grading of HW 2 is done and scores will be out soon
- Solutions to HW 2 and 3

Agenda

1. Announcement
2. Statistical Learning Theory
3. Gradient Descent
4. Regularization
5. SVMs
6. SVMs
7. Kernelization

Statistical Learning Theory - Overview

- Concepts of prediction function, loss function and risk minimization
- Risk estimation and empirical risk minimization
- Error Decomposition

Bayes Prediction Function

If loss function is square loss, then what is the Bayes Predictor?

$$\ell(f(x), y) = (f(x) - y)^2$$

Bayes Prediction Function

If loss function is square loss, then what is the Bayes Predictor?

$$f^*(x) = \arg\min_f R(f) = \arg\min_f E[\ell(f(x), y)]$$

$$= \arg\min_f \int (f(x) - y)^2 p(y|x) dy$$

$$= \arg\min_f \int (f(x)^2 + y^2 - 2f(x)y) p(y|x) dy$$

Taking derivative w.r.t. $f(x)$ set to zero:

$$2(f(x) - E[Y|X]) = 0$$

Hence $f^*(x) = E[Y|X]$.
Reference

Bayes Prediction Function

Similarly:

- If loss function is square loss, then $f^*(x) = E[Y|X = x]$
- If loss function is absolute loss, then $f^*(x)$ is the median of the distribution of Y conditioned on $X = x$. (Exercise)
- If $\mathcal{Y}$ is discrete and loss function is $0 - 1$ loss, then $f^*(x) = \underset{c \in \mathcal{Y}}{argmax}\ p(y = c|x)$. (Lecture 1)

Bayes Prediction Function (Example Question)

Question: Let x be sampled uniformly from $\{-100, -99, \ldots, 99, 100\}$. For every sample x_i, y_i is generated as $y_i = x_i + \eta$, $\eta \sim \mathcal{N}(0, \sigma)$, $\sigma > 0$. What is the Bayes prediction function under L_2 and L_1 loss?

Bayes Prediction Function - Solution

Generating distribution for $y_i \sim \mathcal{N}(x_i, \sigma)$.

- If loss function is $L2$, then $f^*(x) = E[Y|X = x]$. That is the mean, hence $f^*(x) = x$
- If loss function is $L1$, then $f^*(x)$ is the median of the distribution of Y conditioned on $X = x$. As the median of Gaussian distribution is the same as its mean, $f^*(x) = x$

Error Decomposition - I

Select true of false for each of the following statements:
Question If the hypothesis space consists of all possible functions, then approximation error is non-zero.
Recall definition $R(f_{\mathcal{F}}) - R(f^*)$

Error Decomposition - I

Select true of false for each of the following statements:
Question If the hypothesis space consists of all possible functions, then approximation error is non-zero.
False - It has to be zero. Hypothesis space would also include f^* leading to $R(f_{\mathcal{F}}) = R(f^*), R(f_{\mathcal{F}}) - R(f^*) = 0$

Error Decomposition - I

Select true of false for each of the following statements:
Question Estimation Error can be negative.
Recall definition $R(\hat{f}_n) - R(f_{\mathcal{F}})$

Error Decomposition - I

Select true of false for each of the following statements:
Question Estimation Error can be negative.
False - by definition $R(\hat{f}_n)$ can at best be equal to $R(f_{\mathcal{F}})$

Error Decomposition - I

Select true of false for each of the following statements:
Question Optimization Error can be negative.
Recall definition $(R(\tilde{f}_n) - R(\hat{f}_n))$

Error Decomposition - I

Select true of false for each of the following statements:
Question Optimization Error can be negative.
True - Due to randomness of optimization algorithm, solution can converge to a $\tilde{f}_n$ that results in lower risk

Error Decomposition - I

Select true of false for each of the following statements:
Question The empirical risk of the ERM, $\hat{R}(\hat{f})$, is an unbiased estimator of the risk of the ERM $R(\hat{f})$. Does your answer change if it's a $\hat{R}(f)$ where f is independent of training data?

Error Decomposition - I

Select true of false for each of the following statements:
Question The empirical risk of the ERM, $\hat{R}(\hat{f})$, is an unbiased estimator of the risk of the ERM $R(\hat{f})$. Does your answer change if it's a $\hat{R}(f)$ where f is independent of training data?
If $\hat{f}$ is learnt from the training data, the empirical risk of the ERM doesn't depict the true distribution risk. This is why we use a test set to approximate its true risk.

Error Decomposition - II

For each, use $\leq$, $\geq$, or $=$ to determine the relationship between the two quantities, or if the relationship cannot be determined. Throughout assume $\mathcal{F}_1, \mathcal{F}_2$ are hypothesis spaces with $\mathcal{F}_1 \subset \mathcal{F}_2$, and assume we are working with a fixed loss function ℓ.

Question The estimation errors of two decision functions f_1, f_2 that minimize the empirical risk over the same hypothesis space, where f_2 uses 5 extra data points.

Error Decomposition - II

For each, use $\leq$, $\geq$, or $=$ to determine the relationship between the two quantities, or if the relationship cannot be determined. Throughout assume $\mathcal{F}_1, \mathcal{F}_2$ are hypothesis spaces with $\mathcal{F}_1 \subset \mathcal{F}_2$, and assume we are working with a fixed loss function ℓ.

Question The estimation errors of two decision functions f_1, f_2 that minimize the empirical risk over the same hypothesis space, where f_2 uses 5 extra data points.

Answer Roughly speaking, more data is better, so we would tend to expect that f_2 will have lower estimation error $(R(\hat{f}_n) - R(f_{\mathcal{F}}))$. That said, this is not always the case, so the relationship cannot be determined.

Error Decomposition - II

For each, use $\leq$, $\geq$, or $=$ to determine the relationship between the two quantities, or if the relationship cannot be determined. Throughout assume $\mathcal{F}_1, \mathcal{F}_2$ are hypothesis spaces with $\mathcal{F}_1 \subset \mathcal{F}_2$, and assume we are working with a fixed loss function ℓ.

Question The approximation errors of the two decision functions f_1, f_2 that minimize risk with respect to $\mathcal{F}_1, \mathcal{F}_2$, respectively (i.e., $f_1 = f_{\mathcal{F}_1}$ and $f_2 = f_{\mathcal{F}_2}$).

Error Decomposition - II

For each, use $\leq$, $\geq$, or $=$ to determine the relationship between the two quantities, or if the relationship cannot be determined. Throughout assume $\mathcal{F}_1, \mathcal{F}_2$ are hypothesis spaces with $\mathcal{F}_1 \subset \mathcal{F}_2$, and assume we are working with a fixed loss function ℓ.

Question The approximation errors of the two decision functions f_1, f_2 that minimize risk with respect to $\mathcal{F}_1, \mathcal{F}_2$, respectively (i.e., $f_1 = f_{\mathcal{F}_1}$ and $f_2 = f_{\mathcal{F}_2}$).

Answer The approximation error $(R(f_{\mathcal{F}})) - R(f^*)$ of f_1 will be larger $(\geq)$.

Error Decomposition - II

For each, use $\leq$, $\geq$, or $=$ to determine the relationship between the two quantities, or if the relationship cannot be determined. Throughout assume $\mathcal{F}_1, \mathcal{F}_2$ are hypothesis spaces with $\mathcal{F}_1 \subset \mathcal{F}_2$, and assume we are working with a fixed loss function ℓ.

Question The empirical risks of two decision functions f_1, f_2 that minimize the empirical risk over $\mathcal{F}_1, \mathcal{F}_2$, respectively. Both use the same fixed training data. What about the actual risk of these two functions?

Error Decomposition - II

For each, use $\leq$, $\geq$, or $=$ to determine the relationship between the two quantities, or if the relationship cannot be determined. Throughout assume $\mathcal{F}_1, \mathcal{F}_2$ are hypothesis spaces with $\mathcal{F}_1 \subset \mathcal{F}_2$, and assume we are working with a fixed loss function ℓ.

Question The empirical risks of two decision functions f_1, f_2 that minimize the empirical risk over $\mathcal{F}_1, \mathcal{F}_2$, respectively. Both use the same fixed training data. What about the actual risk of these two functions?

Answer The empirical risk of f_1 will be larger (specifically $\geq$). We cannot determine the relationship for actual risk.

Gradient Descent - Overview

- How do you solve the optimization problem of ERM? Gradient Descent
- What about if GD is too expensive? SGD
- Trade off noisy optimization for speed of calculation
- Loss Functions

Gradient Descent - I

Decide whether the following statements apply to full batch gradient descent (GD), mini-batch GD, neither, or both.

Assume we're minimizing a differentiable, convex objective function $J(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$, and we are currently at w_t, which is not a minimum. For full batch GD, take $v = \nabla_w J(w_t)$, and for minibatch GD take v to be a mini-batch estimate of $\nabla_w J(w_t)$ based on a random sample of the training data.

Question For any step size $\eta > 0$, after applying the update rule $w_{t+1} \leftarrow w_t - \eta v$. we must have $J(w_{t+1}) < J(w_t)$.

Gradient Descent - I

Decide whether the following statements apply to full batch gradient descent (GD), mini-batch GD, neither, or both.

Assume we're minimizing a differentiable, convex objective function $J(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$, and we are currently at w_t, which is not a minimum. For full batch GD, take $v = \nabla_w J(w_t)$, and for minibatch GD take v to be a mini-batch estimate of $\nabla_w J(w_t)$ based on a random sample of the training data.

Question For any step size $\eta > 0$, after applying the update rule $w_{t+1} \leftarrow w_t - \eta v$. we must have $J(w_{t+1}) < J(w_t)$.

Answer Neither.

- Depends on whether the learning rate is good.
- Moreover, for mini-batch GD, it also depends on whether v is representative enough.

Gradient Descent - II

Decide whether the following statements apply to full batch gradient descent (GD), mini-batch GD, neither, or both.

Assume we're minimizing a differentiable, convex objective function $J(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$, and we are currently at w_t, which is not a minimum. For full batch GD, take $v = \nabla_w J(w_t)$, and for minibatch GD take v to be a mini-batch estimate of $\nabla_w J(w_t)$ based on a random sample of the training data.

Question There must exist some $\eta > 0$ such that after applying the update rule $w_{t+1} \leftarrow w_t - \eta v$ we have $J(w_{t+1}) < J(w_t)$.

Gradient Descent - II

Decide whether the following statements apply to full batch gradient descent (GD), mini-batch GD, neither, or both.

Assume we're minimizing a differentiable, convex objective function $J(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$, and we are currently at w_t, which is not a minimum. For full batch GD, take $v = \nabla_w J(w_t)$, and for minibatch GD take v to be a mini-batch estimate of $\nabla_w J(w_t)$ based on a random sample of the training data.

Question There must exist some $\eta > 0$ such that after applying the update rule $w_{t+1} \leftarrow w_t - \eta v$ we have $J(w_{t+1}) < J(w_t)$.

Answer True for full batch. For mini-batch GD, it depends on whether v is representative enough.

Regularization - Overview

- During optimization, we might encounter the trade-off between approximation error and estimation error $\rightarrow$ Regularization
- L1 vs L2 Regularization
  - Feature selection through L1
- Coordinate descent for L1

Regularization - Question

We solve lasso and ridge regression where input lives in $\mathcal{R}^4$. The first two features of all the input vector are duplicates of each other, or $x_{i1} = x_{i2}$ for all i. Consider the following weight vectors:

1. $(0, 1.2, 6.7, 2.1)^T$
2. $(0.6, 0.6, 6.7, 2.1)^T$
3. $(1.2, 0, 6.7, 2.1)^T$
4. $(-0.1, 1.3, 6.7, 2.1)^T$

Which of them are valid solution for a) Ridge Regression and b) Lasso Regression?

Regularization - Solution

a) Ridge Regression

2 $(0.6, 0.6, 6.7, 2.1)^T$ - ℓ_2 regularization spreads weight evenly for identical features

b) Lasso Regression

1,2,3 - ℓ_1 regularization spreads weight arbitrarily (all weights same sign)

SVMs - Overview

- Existence of multiple candidate hyperplane $\rightarrow$ SVMs/Margin Maximization
- Hard and Soft Margin SVMs
- Subgradient descent to solve the primal optimization

SVMs - Overview

- Existence of multiple candidate hyperplane $\rightarrow$ SVMs/Margin Maximization
- Hard and Soft Margin SVMs
- Subgradient descent to solve the primal optimization
- We also have strong duality, so the solution to the dual (and corresponding primal optimum) has interesting properties $\rightarrow$ Complementary Slackness
- Complementary slackness $\rightarrow$ Dependence of the solution on only a few "support vectors"

SVMs - I

If we fit the data in Fig. 1 using a hard margin SVM, then what are the support vectors? (Colours correspond to labels)

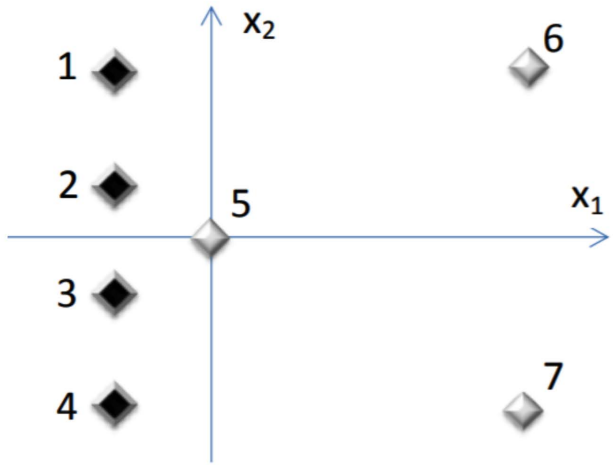

Figure: Train Data

SVMs - I

If we fit the data in Fig. 1 using a hard margin SVM, then what are the support vectors? (Colours correspond to labels)
Hard Margin $\rightarrow$ Have to separate the points $\rightarrow$ Support vectors are 1,2,3,4,5

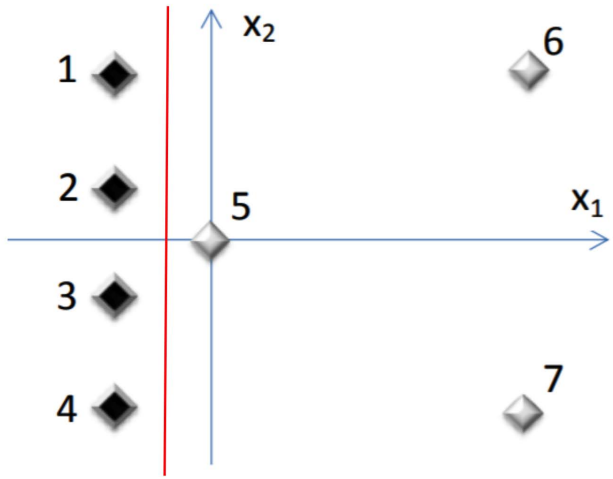

Figure: Train Data with a margin

SVMs - II

If you could remove one point on Fig. 1 to allow for a large margin using a hard margin SVM, which point is it? (Alternately, if you had a soft margin SVM then what point would likely have associated slack)

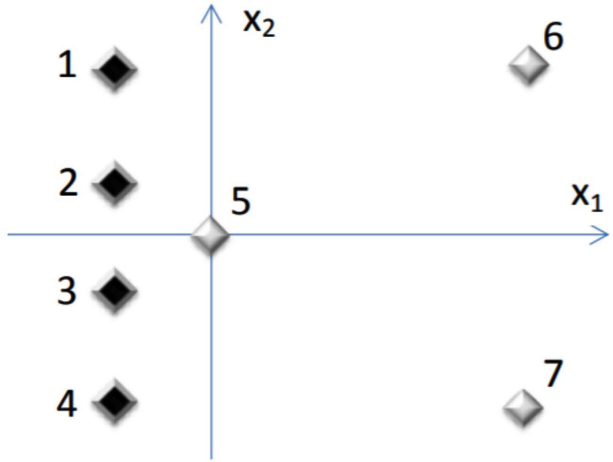

Figure: Train Data

SVMs - II

If you could remove one point on Fig. 1 to allow for a large margin using a hard margin SVM, which point is it? (Alternately, if you had a soft margin SVM then what point would likely have associated slack)
Point 5 because it makes a big difference to the margin associated with the chosen hyperplane

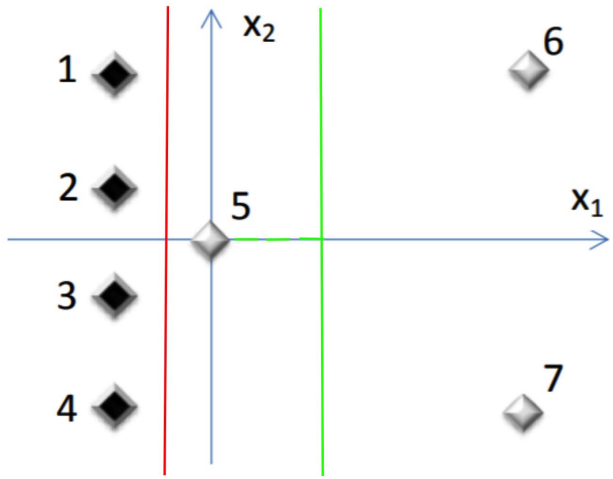

Figure: Train Data

Kernelization - Overview

- Linear features might not suffice. We might want interaction/transformation of the features
- Writing the SVM in its dual form, we see that it can be kernelized (feature vector only appears as an inner product)
- Swap this inner product with a kernel
- Solution in the span of the data $\rightarrow$ Representer theorem

Kernelization - I

Consider the objective function

$$J(w) = \|Xw - y\|_1 + \lambda \|w\|_2^2$$

Assume we have a positive semidefinite kernel k.

1. What is the kernelized version of this objective?
2. Given a new test point x, find the predicted value.

Kernelization - I

1. $J(\alpha) = \|K\alpha - y\|_1 + \lambda \alpha^T K \alpha$, where $K_{ij} = k(x_i, x_j)$. Here x_i^T is the ith row of X. (Lecture 5/Recitation 5)
2. $f_\alpha(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$.

Kernelization - II

Consider the following dataset, where each point is an example in $\mathbb{R}^2$. Can you get 100% training accuracy with a linear classifier? Suggest a new feature that will allow you to make the data separable.

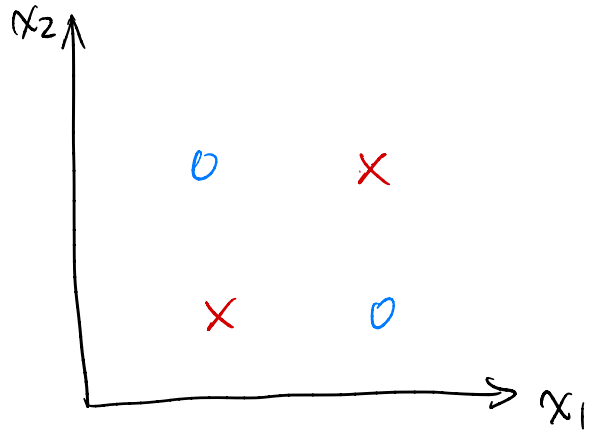

Figure: Train dataset

Kernelization - II

Consider the following dataset, where each point is an example in $\mathbb{R}^2$. Can you get 100% training accuracy with a linear classifier? Suggest a new feature that will allow you to make the data separable.
No, classic XOR case. But we if add $(x_1 - x_2)^2$ as a feature then the data becomes linearly separable (many possible answers here - polynomial kernels, RBF kernels and so on)

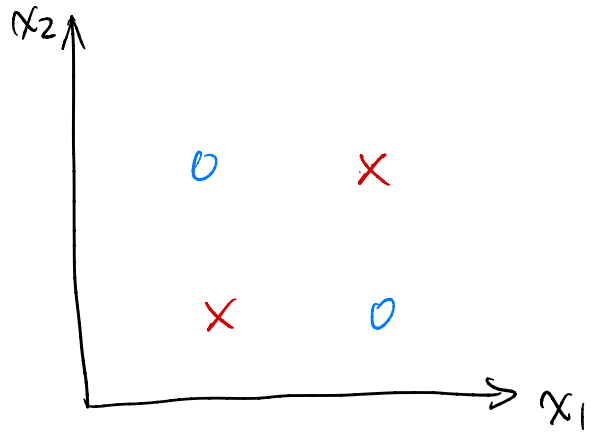

Figure: Train dataset

Good luck!

- Prepare fundamentals from lectures/homework
- Past exams/solutions to get a flavour of the questions
- Expected to be shorter than past years but budget time to upload solutions