Midterm Review Solution

DS-GA 1003 Machine Learning

NYU CDS

March 22, 2021

Contents

Learning Theory Framework

Bayes Prediction Function

- If loss function is $L2$, then $f^*(x) = E[Y|X = x]$
- if loss function is $L1$, then $f^*(x)$ is the median of the distribution of Y conditioned on $X = x$.
- If $\mathcal{Y}$ is discrete and loss function is $0-1$ loss, then $f^*(x) = \underset{c \in \mathcal{Y}}{\textit{argmax}}\ p(y = c|x)$

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:

1. Approximation Error is a Random Variable
2. Estimation Error is a Random Variable
3. Optimization Error is a Random Variable.
4. If the hypothesis space consists of all possible functions, then approximation error is non-zero.
5. Estimation Error can be negative.
6. Optimization Error can be negative.
7. 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?

Solution

1. **False** - Approximation Error (of $\mathcal{F}$) $= R(f_{\mathcal{F}}) - R(f^*)$, where risk $R(f) = \mathbb{E}\ell(f(X), y)$ - is a deterministic quantity
2. **True** - Estimation error (of $\hat{f}_n$ in $\mathcal{F}$) $= R(\hat{f}_n) - R(f_{\mathcal{F}})$, where $\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$ is dependent on random sample of size n
3. **True** - Optimization Error (of $\tilde{f}_n$) $= R(\tilde{f}_n) - R(\hat{f}_n)$, where $\tilde{f}_n$ is the function our optimization method returns - also dependent of random data sample
4. **False** - It would be zero. Hypothesis space would also include f^* leading to $R(f_{\mathcal{F}}) = R(f^*)$
5. **False** - by definition above $R(\hat{f}_n)$ can at best be equal to $R(f_{\mathcal{F}})$
6. **True** - Due to randomness of optimization algorithm, solution can converge to a $\tilde{f}_n$ that results in lower risk
7. 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.
   - **False**
   - **True**

Error Decomposition - II

For each, use $\leqslant$, $\geqslant$, 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.

1. 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.
2. 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 $f_2 = f_{\mathcal{F}_2}$).
3. 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.
4. The estimation errors (for $\mathcal{F}_1, \mathcal{F}_2$, respectively) of two decision functions f_1, f_2 that minimize the empirical risk over $\mathcal{F}_1, \mathcal{F}_2$, respectively.
5. The risk of two decision functions f_1, f_2 that minimize the empirical risk over $\mathcal{F}_1, \mathcal{F}_2$, respectively.

Solution

1. Roughly speaking, more data is better, so we would tend to expect that f_2 will have lower estimation error. That said, this is not always the case, so the relationship cannot be determined.
2. The approximation error of f_1 will be larger.
3. The empirical risk of f_1 will be larger.
4. Roughly speaking, increasing the hypothesis space should increase the estimation error since the approximation error will decrease, and we expect to need more data. That said, this is not always the case, so the answer is the relationship cannot be determined.
5. Cannot be determined.

Regularization

Correlated features

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?

Correlated features - 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)

Optimization

Question on Subgradient

Definition (Subgradient and Subdifferential)

A vector g is a subgradient of (convex) $f : \mathcal{R}^d \to \mathcal{R}$ at x if for all z

$$f(z) \geqslant f(x) + g^T(z - x)$$

. The set of all subgradients at x is called the subdifferential of f at x $\partial f(x)$

Questions:

1. (True/False) If f is convex and differentiable everywhere in the domain, then $\partial f(x) = \{\nabla f(x)\}$
2. (True/False) The subdifferential of f at x, $\partial f(x)$ is always a convex set. (Null set is trivially complex)

Subgradient - Solution

1. **(True) If f is convex and differentiable everywhere in the domain, then $\partial f(x) = \{\nabla f(x)\}$**
   By the gradient (first-order) conditions for convexity, we know that $\nabla f(x) \in \partial f(x)$. Next suppose $g \in \partial f(x)$. This means that for all $v \in \mathbb{R}^n$ and $h \in \mathbb{R}$ we have:

$$f(x+hv) \geqslant f(x) + hg^T v \Longrightarrow \frac{f(x+hv)-f(x)}{h} \geqslant g^T v$$

   Using $-h$ in place of h gives

$$f(x-hv) \geqslant f(x) - hg^T v \Longrightarrow g^T v \geqslant \frac{f(x-hv)-f(x)}{-h}$$

   Taking limits as $h \to 0$ gives $\nabla f(x)^T v \geqslant g^T v \geqslant \nabla f(x)^T v$
   Thus all terms are equal. Subtracting gives $(\nabla f(x) - g)^T v = 0$ which holds for all $v \in \mathbb{R}^n$.
   Letting $v = \nabla f(x) - g$ proves $\|\nabla f(x) - g\|_2^2 = 0$ giving the result.

Subgradient - Solution

(True) The subdifferential of f at x, $\partial f(x)$ is always a convex set.
Fix $f : \mathbb{R}^n \to \mathbb{R}$ and $x \in \mathbb{R}^n$. Then the subdifferential $\partial f(x)$ is a convex set.
Let $g_1, g_2 \in \partial f(x)$ and $t \in (0,1)$. We must show $(1-t)g_1 + tg_2$ is a subgradient. Note that, for any $y \in \mathbb{R}^n$, we have

$$\begin{aligned} f(x) + ((1-t)g_1 + tg_2)^T (y-x) &= (1-t)\left(f(x) + g_1^T(y-x)\right) + t\left(f(x) + g_2^T(y-x)\right) \\ &\leqslant (1-t)f(y) + tf(y) \\ &= f(y) \end{aligned}$$

Question on Gradient Descent

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.

1. For any step size $\eta > 0$, after applying the update rule $w_{t+1} \leftarrow w_t - \eta v$. we must have $J(w_{t+1}) < J(w_t)$.
2. There must exist some $\eta > 0$ such that after applying the update rule $w_{t+1} \leftarrow w_t - \eta v$ we have $J(w_{t+1}) < J(w_t)$.
3. v is an unbiased estimator of the full batch gradient.

Gradient Descent - Solution

1. Neither.
   - Depends on whether the learning rate is good.
   - Moreover, for mini-batch GD, it also depends on whether v is representative enough.
2. Full batch.
   - For mini-batch GD, it depends on whether v is representative enough.
3. Both.
   - Proved in lecture

Classification

Question on Classification

Suppose $x_1, \ldots, x_n \in \mathbb{R}^d$ and $y_1, \ldots, y_n \in \{-1, 1\}$. Here we look at y_i as the label of x_i. We say the data points are linearly separable if there is a vector $v \in \mathbb{R}^d$ and $a \in \mathbb{R}$ such that $v^T x_i > a$ when $y_i = 1$ and $v^T x_i < a$ for $y_i = -1$. Give a method for determining if the given data points are linearly separable.

Solution

- Solve the hard-margin SVM problem:

$$\begin{aligned} \text{minimize}_{w,b} \quad & \|w\|_2^2 \\ \text{subject to} \quad & y_i \left(w^T x_i + b \right) \geqslant 1; \text{ for all } i = 1, \ldots, n \end{aligned}$$

- If the resulting problem is feasible, then the data is linearly separable

The Representer Theorem and Kernelization

Kernelization

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 - Solution

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.
2. $f_\alpha(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$.

MLE and Conditional Probability Models

Maximum Likelihood Estimation

1. Suppose we have samples $x_1, \ldots, x_n$ i.i.d. drawn from uniform distribution $\mathcal{U}(-a, a)$. Find the maximum likelihood estimator of a.
2. Which of the following models can be learned by MLE?
   - Perceptron
   - Logistic regression
   - SVM

Maximum Likelihood Estimation - Solution

1.
   - The likelihood is:
   $$L(-a, a) = \Pi_{i=1}^{n} \left(\frac{1}{2a} 1_{[-a,a]}(x_i) \right)$$
   - The likelihood is greater than zero if and only $-a \leqslant min(x_1, \ldots, x_n)$ and $a \geqslant max(x_1, \ldots, x_n)$.
   - When above conditions are satisfied, the likelihood is a monotonically decreasing function of $2a$.
   - And the smallest a will be attained when $a = max(|x_1|, \ldots, |x_n|)$ to satisfy the conditions.
   - Therefore, $a = max(|x_1|, \ldots, |x_n|)$ give us the MLE.

2. Logistic Regression
   - Only probabilistic model amongst the three, utilizes Bernoulli distribution