CSC2515 Midterm Review
Part 1

Sicong Huang Haoran Zhang

Oct 16, 2020

Midterm Information

What might be on the midterm?

- Everything covered in detail during lecture

What will NOT be on the midterm?

- Programming
- New concepts introduced in Homeworks (Ex: KL divergence, Huber loss)
- Anything from tutorials that wasn't in the lectures slides (Ex: SVD, Shannon's source coding theorem)
- Anything from the textbooks that wasn't in the lecture slides

Midterm Topics (Today)

- Lecture 1
  - k-Nearest Neighbors
  - Bayes Optimality
- Lecture 2
  - Decision trees and Information theory (information gain)
  - Bias Variance
  - Bagging
- Lecture 3
  - Linear regression
  - Logistic regression

Midterm Topics (Today)

- Lecture 4
  - Gradient descent
  - L^1,L^2 regularization (covered in previous tutorial, see Q3 in 19 midterm)
  - SVMs (covered in previous tutorial, see Q6 in 19 midterm)
  - Boosting, additive models

Midterm Topics (Next Time)

- Lecture 5
  - PCA
  - Linear and non-linear autoencoders
  - K-means
  - Maximum likelihood estimation (MLE)
- Lecture 6
  - Full Bayesian parameter estimation
  - Maximum a-posteriori (MAP)
  - Naive Bayes
  - Gaussian discriminant analysis

KNN (19 Fall midterm Q7)

- When we analyzed KNN, we assumed the training examples were sampled densely enough so that the true conditional probability $p(t \mid \mathrm{x})$ is approximately constant in the vicinity of a query point $\mathrm{x}_\star$. Suppose it is a binary classification task with targets $t \in \{0, 1\}$ and $p(t = 1 \mid \mathrm{x}_\star) = 0.6$.
- What is the asymptotic error rate at $\mathrm{x}_\star$ for a 1-nearest-neighbor classifier? (By asymptotic, I mean as the number of training examples $N \to \infty$.) Justify your answer.

KNN (19 Fall midterm Q7)

- When we analyzed KNN, we assumed the training examples were sampled densely enough so that the true conditional probability $p(t \mid \mathrm{x})$ is approximately constant in the vicinity of a query point $\mathrm{x}_\star$. Suppose it is a binary classification task with targets $t \in \{0, 1\}$ and $p(t = 1 \mid \mathrm{x}_\star) = 0.6$.
- What is the asymptotic error rate at $\mathrm{x}_\star$ for a 1-nearest-neighbor classifier? (By asymptotic, I mean as the number of training examples $N \to \infty$.) Justify your answer.

  Let $t_\star$ denote the true target and t_N denote the target at the nearest neighbor. These are independent Bernoulli random variables with parameter 0.6. The classifier makes a mistake if $t_\star = 0$ and $t_N = 1$ or if $t_\star = 1$ and $t_N = 0$. Hence, the probability of a mistake, i.e. the error rate, is $0.4 \cdot 0.6 + 0.6 \cdot 0.4 = 0.48$.

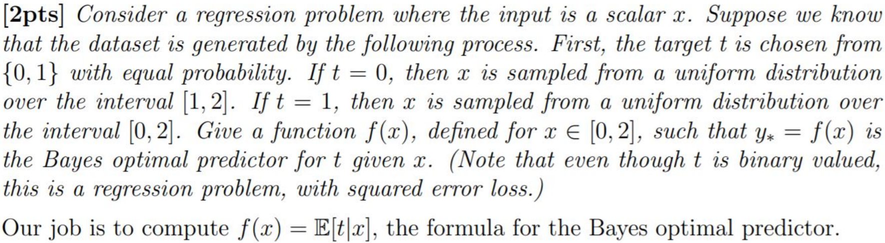
KNN (19 Fall midterm Q7)

- When we analyzed KNN, we assumed the training examples were sampled densely enough so that the true conditional probability $p(t \mid \mathrm{x})$ is approximately constant in the vicinity of a query point $\mathrm{x}_\star$. Suppose it is a binary classification task with targets $t \in \{0, 1\}$ and $p(t = 1 \mid \mathrm{x}_\star) = 0.6$.
- Approximately what is the asymptotic (as $N \to \infty$) error rate at $\mathrm{x}_\star$ for a K-nearest-neighbors classifier when K is very large? Justify your answer.

KNN (19 Fall midterm Q7)

- When we analyzed KNN, we assumed the training examples were sampled densely enough so that the true conditional probability $p(t \mid \mathrm{x})$ is approximately constant in the vicinity of a query point $\mathrm{x}_\star$. Suppose it is a binary classification task with targets $t \in \{0, 1\}$ and $p(t = 1 \mid \mathrm{x}_\star) = 0.6$.
  For large K, the asymptotic KNN error rate is approximately the Bayes error rate. In this example, the Bayes classifier will predict $y = 1$. Hence, the error rate is 0.4.

Bayes Optimality (18 Fall midterm A Q9)



[2pts] *Consider a regression problem where the input is a scalar x. Suppose we know that the dataset is generated by the following process. First, the target t is chosen from $\{0, 1\}$ with equal probability. If $t = 0$, then x is sampled from a uniform distribution over the interval $[1, 2]$. If $t = 1$, then x is sampled from a uniform distribution over the interval $[0, 2]$. Give a function $f(x)$, defined for $x \in [0, 2]$, such that $y_* = f(x)$ is the Bayes optimal predictor for t given x. (Note that even though t is binary valued, this is a regression problem, with squared error loss.)*

Our job is to compute $f(x) = \mathbb{E}[t|x]$, the formula for the Bayes optimal predictor.

Bayes Optimality (18 Fall midterm A Q9)

Note that $p(x|t=1) = 1/2$ on $[0,2]$ and 0 elsewhere, and $p(x|t=0) = 1$ on $[1,2]$ and 0 elsewhere. By Bayes' Rule, for $x \in [0,1]$,

$$\begin{aligned} p(t=1|x) &= \frac{p(t=1)p(x|t=1)}{p(t=0)p(x|t=0) + p(t=1)p(x|t=1)} \\ &= \frac{\frac{1}{2}\cdot\frac{1}{2}}{\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot\frac{1}{2}} \\ &= 1 \end{aligned}$$

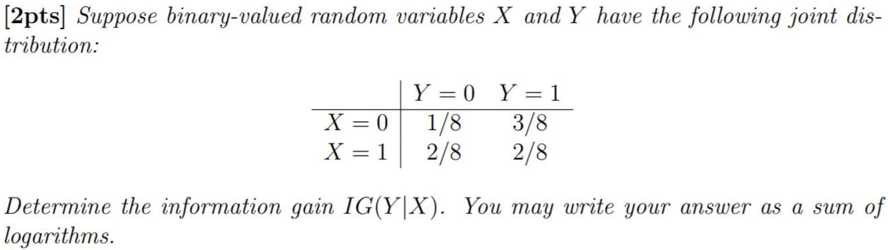
And for $x \in [1,2]$,

$$\begin{aligned} p(t=1|x) &= \frac{p(t=1)p(x|t=1)}{p(t=0)p(x|t=0) + p(t=1)p(x|t=1)} \\ &= \frac{\frac{1}{2}\cdot\frac{1}{2}}{\frac{1}{2}\cdot 1 + \frac{1}{2}\cdot\frac{1}{2}} \\ &= \frac{1}{3} \end{aligned}$$

Hence,

$$f(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ \frac{1}{3} & \text{if } 1 < x \le 2 \end{cases}.$$

Information theory (18 Fall midterm B Q7)

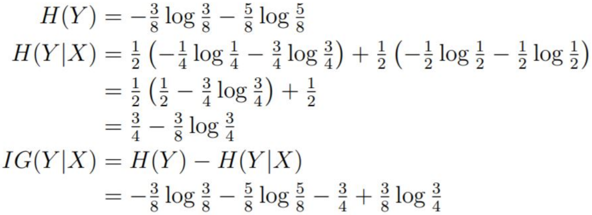


[2pts] *Suppose binary-valued random variables X and Y have the following joint distribution:*

	$Y=0$	$Y=1$
$X=0$	$1/8$	$3/8$
$X=1$	$2/8$	$2/8$

Determine the information gain $IG(Y|X)$. You may write your answer as a sum of logarithms.

Information theory (18 Fall midterm B Q7)



$$\begin{aligned}
H(Y) &= -\tfrac{3}{8}\log\tfrac{3}{8} - \tfrac{5}{8}\log\tfrac{5}{8} \\
H(Y|X) &= \tfrac{1}{2}\left(-\tfrac{1}{4}\log\tfrac{1}{4} - \tfrac{3}{4}\log\tfrac{3}{4}\right) + \tfrac{1}{2}\left(-\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{2}\log\tfrac{1}{2}\right) \\
&= \tfrac{1}{2}\left(\tfrac{1}{2} - \tfrac{3}{4}\log\tfrac{3}{4}\right) + \tfrac{1}{2} \\
&= \tfrac{3}{4} - \tfrac{3}{8}\log\tfrac{3}{4} \\
IG(Y|X) &= H(Y) - H(Y|X) \\
&= -\tfrac{3}{8}\log\tfrac{3}{8} - \tfrac{5}{8}\log\tfrac{5}{8} - \tfrac{3}{4} + \tfrac{3}{8}\log\tfrac{3}{4}
\end{aligned}$$

Bias Variance (modified from CSC2515 19 midterm Q4)

- Carol and Dave are each trying to predict stock prices using neural networks. They formulate this as a regression problem using squared error loss. Carol trains a single logistic regression model on a certain training set and uses its predictions on the test set. Dave trains 5 different models (using exactly the same architecture, training data, etc. as Carol) starting with different random initializations, and averages their predictions on the test set.

  For each of the following questions, please briefly and informally justify your answer. You do not need to provide a mathematical proof.
- Compared with Carol's approach, is the Bayes error for Dave's approach HIGHER, LOWER, or THE SAME?

Bias Variance (modified from CSC2515 19 midterm Q4)

- Carol and Dave are each trying to predict stock prices using neural networks. They formulate this as a regression problem using squared error loss. Carol trains a single logistic regression model on a certain training set and uses its predictions on the test set. Dave trains 5 different models (using exactly the same architecture, training data, etc. as Carol) starting with different random initializations, and averages their predictions on the test set.

  For each of the following questions, please briefly and informally justify your answer. You do not need to provide a mathematical proof.
- [4 points] Compared with Carol's approach, is the Bayes error for Dave's approach HIGHER, LOWER, or THE SAME?

  THE SAME. The Bayes error is a property of the data generating distribution, and doesn't depend on the algorithm that was used.

Bias Variance (modified from CSC2515 19 midterm Q4)

- Carol and Dave are each trying to predict stock prices using neural networks. They formulate this as a regression problem using squared error loss. Carol trains a single logistic regression model on a certain training set and uses its predictions on the test set. Dave trains 5 different models (using exactly the same architecture, training data, etc. as Carol) starting with different random initializations, and averages their predictions on the test set.
  For each of the following questions, please briefly and informally justify your answer. You do not need to provide a mathematical proof.
- Compared with Carol's approach, is the bias for Dave's approach HIGHER, LOWER, or THE SAME?

Bias Variance (modified from CSC2515 19 midterm Q4)

- Carol and Dave are each trying to predict stock prices using neural networks. They formulate this as a regression problem using squared error loss. Carol trains a single logistic regression model on a certain training set and uses its predictions on the test set. Dave trains 5 different models (using exactly the same architecture, training data, etc. as Carol) starting with different random initializations, and averages their predictions on the test set.

  For each of the following questions, please briefly and informally justify your answer. You do not need to provide a mathematical proof.
- Compared with Carol's approach, is the bias for Dave's approach HIGHER, LOWER, or THE SAME?

  THE SAME. Sampling multiple hypotheses from the same distribution and averaging their predictions doesn't change the expected predictions due to linearity of expectation. Hence it doesn't change the bias.

Bias Variance (modified from CSC2515 19 midterm Q4)

- Carol and Dave are each trying to predict stock prices using neural networks. They formulate this as a regression problem using squared error loss. Carol trains a single logistic regression model on a certain training set and uses its predictions on the test set. Dave trains 5 different models (using exactly the same architecture, training data, etc. as Carol) starting with different random initializations, and averages their predictions on the test set.

  For each of the following questions, please briefly and informally justify your answer. You do not need to provide a mathematical proof.
- Compared with Carol's approach, is the variance for Dave's approach HIGHER, LOWER, or THE SAME?

Bias Variance (modified from CSC2515 19 midterm Q4)

- Compared with Carol's approach, is the variance for Dave's approach HIGHER, LOWER, or THE SAME?

  LOWER. Averaging over multiple samples reduces the variance of the predictions, even if those samples are not fully independent. (In this case, they're not fully independent as they share the same training set.)

  Marking: each part was worth 2 points for the correct answer and 2 points for the explanation. For part (b), mentioning linearity of expectation or some other argument involving expected value was required for full marks. Note that the procedure here isn't bagging, so answers involving bagging lost some points.

Regression (19 midterm review Q2)

Given input $\mathbf{x} \in \mathbb{R}^d$ and target $y \in \mathbb{R}$, define $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\epsilon}$ to be a noisy pertubation of $\mathbf{x}$ where we assume

- $\mathbb{E}[\epsilon_i] = 0$
- for $i \neq j$: $\mathbb{E}[\epsilon_i \epsilon_j] = 0$
- $\mathbb{E}[\epsilon_i^2] = \lambda$

We define the following objective that tries to be robust to noise

$$\mathbf{w}^* = \arg\min \mathbb{E}_\epsilon[(\mathbf{w}^T \hat{\mathbf{x}} - y)^2] \tag{1}$$

Show that it is equivalent to minimizing L_2 regularized linear regression, i.e.

$$\mathbf{w}^* = \arg\min \left[(\mathbf{w}^T \mathbf{x} - y)^2 + \lambda ||\mathbf{w}||^2 \right] \tag{2}$$

Regression (19 midterm review Q2)

We can write the inner term as,

$$\begin{aligned}
(\mathbf{w}^T\hat{\mathbf{x}} - y)^2 &= (\mathbf{w}^T\mathbf{x} + \mathbf{w}^T\epsilon - y)^2 \tag{3}\\
&= (\mathbf{w}^T\mathbf{x} - y)^2 + 2\mathbf{w}^T\epsilon(\mathbf{w}^T\mathbf{x} - y) + (\mathbf{w}^T\epsilon)^2 \tag{4}\\
&= (\mathbf{w}^T\mathbf{x} - y)^2 + 2\mathbf{w}^T\epsilon(\mathbf{w}^T\mathbf{x} - y) + (\mathbf{w}^T\epsilon^T\epsilon\mathbf{w}) \tag{5}
\end{aligned}$$

Under the expectation the second term will be zero as it is a linear combination of the elements of ϵ. The final term will be the quadratic form of $\mathbf{w}$ with the covariance of ϵ. The covariance is simply λI. Thus we are minimizing,

$$(\mathbf{w}^T\mathbf{x} - y)^2 + \lambda||\mathbf{w}||^2$$

which is exactly the objective of L2-regularized linear regression.

SVM + Gradients (18 Fall midterm B Q9)

[2pts] Recall that the soft-margin SVM can be viewed as minimizing the hinge loss with an L_2 regularization term. I.e.,

$$\begin{aligned} z &= \mathbf{w}^\top \mathbf{x} + b \\ \mathcal{L}(z, t) &= \max(0, 1 - tz) \\ \mathcal{J}(\mathbf{w}, b) &= \tfrac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(z^{(i)}, t^{(i)}). \end{aligned}$$

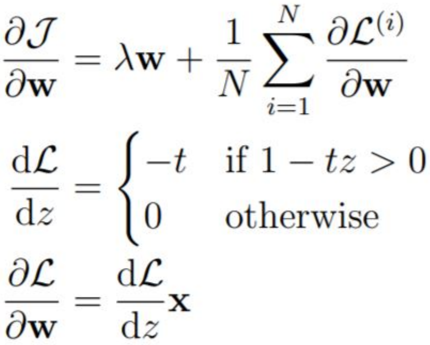
Here, $t \in \{-1, +1\}$. Complete the formulas for the gradient calculations. You don't need to show your work.

$$\frac{\partial \mathcal{J}}{\partial \mathbf{w}} = \underline{\hspace{6cm}} + \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \mathcal{L}^{(i)}}{\partial \mathbf{w}} \qquad \text{(fill in the blank)}$$

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}z} =$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \qquad\qquad \text{(give in terms of } \frac{\mathrm{d}\mathcal{L}}{\mathrm{d}z}\text{)}$$

SVM + Gradients (18 Fall midterm B Q9)



$$\frac{\partial \mathcal{J}}{\partial \mathbf{w}} = \lambda \mathbf{w} + \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathcal{L}^{(i)}}{\partial \mathbf{w}}$$

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}z} = \begin{cases} -t & \text{if } 1 - tz > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\mathrm{d}\mathcal{L}}{\mathrm{d}z} \mathbf{x}$$