18-661: Introduction to ML for Engineers

Mini Exam 1 Review

Spring 2025

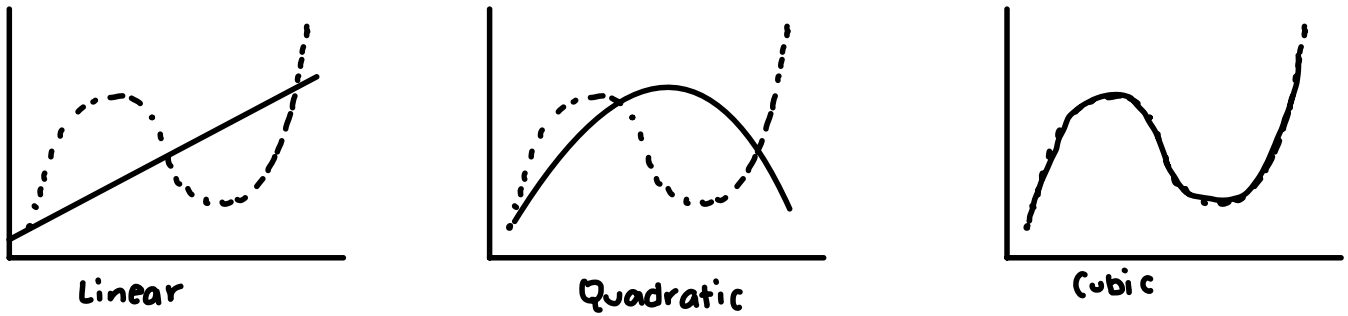
ECE – Carnegie Mellon University

Gradescope Participation Quiz 02/03

Question 1: Suppose we are given a dataset (x_1, x_n) for $n = 1 \ldots N$ where each x_i, y_i are scalars. We use polynomial regression to fit a function $f(x) = w_0 + w_1 x + \cdots + w_M x^M$ to the data. As we increase the degree of the polynomial M we get a model with -

A) Higher bias, lower variance

B) Lower bias, higher variance



Gradescope Participation Quiz 02/03

Question 1: Suppose we are given a dataset (x_1, x_n) for $n = 1 \dots N$ where each x_i, y_i are scalars. We use polynomial regression to fit a function $f(x) = w_0 + w_1 x + \cdots + w_M x^M$ to the data. As we increase the degree of the polynomial M we get a model with -

A) Higher bias, lower variance

B) Lower bias, higher variance

Solution: B

Explanation

Increasing the degree of the polynomial M is equivalent to increasing the model complexity. This will likely lead to lower training loss (AKA lower bias). However, this more complicated model may overfit to the training dataset which means that smaller changes in the training dataset will lead to larger changes in the model

Gradescope Participation Quiz 02/03

Question 2: Which of the following statements are true about addition of the regularizer term $\lambda||\mathbf{w}||^2$ to the residual sum of squares objective function -

A) The solution $\mathbf{w}$ with $\lambda > 0$ will have a higher training loss than the solution with $\lambda = 0$

True, reg. worsens training loss to improve generalization

B) The solution $\mathbf{w}$ with $\lambda > 0$ will have a lower training loss than the solution with $\lambda = 0$

C) The solution $\mathbf{w}$ with $\lambda > 0$ will have a larger L_2 norm than the solution with $\lambda = 0$

D) The solution $\mathbf{w}$ with $\lambda > 0$ will have a smaller L_2 norm than the solution with $\lambda = 0$

True, reg. punishes large magnitudes of w in the loss function leading to smaller value of ||w||

Gradescope Participation Quiz 02/03

Question 2: Which of the following statements are true about addition of the regularizer term $\lambda||\mathbf{w}||^2$ to the residual sum of squares objective function -

A) The solution $\mathbf{w}$ with $\lambda > 0$ will have a higher training loss than the solution with $\lambda = 0$

B) The solution $\mathbf{w}$ with $\lambda > 0$ will have a lower training loss than the solution with $\lambda = 0$

C) The solution $\mathbf{w}$ with $\lambda > 0$ will have a larger L_2 norm than the solution with $\lambda = 0$

D) The solution $\mathbf{w}$ with $\lambda > 0$ will have a smaller L_2 norm than the solution with $\lambda = 0$

Solution: A, D

Gradescope Participation Quiz 02/05

Question 1: Suppose you use a simple bag-of-words Naive Bayes model to determine whether a document was written by an AI generator or a human. Given a set of documents as training data, you find that the prior probability of the "AI generator" class, $\pi_{\text{AI generator}}$, is 0.7, and that the conditional probability of the word "machine" given this class, $P(\textit{machine} \mid \text{class}=\text{"AI generator"})$, is 0.02.

Now suppose you add five new documents to your training data, each of which was written by an AI generator. You re-compute the Naive Bayes model parameters on your training data with these additional documents. Which of the following are feasible updated values of $\pi_{\text{AI generator}}$ and $P(\textit{machine} \mid \text{class}=\text{"AI generator"})$*.

A) $\pi_{\text{AI generator}} = 0.5$, $P(\textit{machine} \mid \text{class}=\text{"AI generator"}) = 0.2$

✓ B) $\pi_{\text{AI generator}} = 0.75$, $P(\textit{machine} \mid \text{class}=\text{"AI generator"}) = 0.3$

C) $\pi_{\text{AI generator}} = 0.7$, $P(\textit{machine} \mid \text{class}=\text{"AI generator"}) = 0.3$

D) $\pi_{\text{AI generator}} = 0.8$, $P(\textit{machine} \mid \text{class}=\text{"AI generator"}) = 0$

*Note: Naive Bayes/Logistic Regression will **not** be on the mini-quiz

Gradescope Participation Quiz 02/05

Question 1: Which of the following are feasible updated values of $\pi_{\text{AI generator}}$ and $P(\textit{machine} \mid \text{class}=\text{“AI generator”})$.

A) $\pi_{\text{AI generator}} = 0.5$, $P(\textit{machine} \mid \text{class}=\text{“AI generator”}) = 0.2$

B) $\pi_{\text{AI generator}} = 0.75$, $P(\textit{machine} \mid \text{class}=\text{“AI generator”}) = 0.3$

C) $\pi_{\text{AI generator}} = 0.7$, $P(\textit{machine} \mid \text{class}=\text{“AI generator”}) = 0.3$

D) $\pi_{\text{AI generator}} = 0.8$, $P(\textit{machine} \mid \text{class}=\text{“AI generator”}) = 0$

Solution: B

Explanation

$\pi_{\text{AI generator}}$ should increase since you have more occurrences of AI generated documents. $P(\textit{machine} \mid \text{class}=\text{“AI generator”})$ cannot become 0.

Gradescope Participation Quiz 02/05

Question 2: Consider a binary classification problem. Suppose that you train a naive Bayes model to solve this problem, using Laplacian smoothing with parameter α to compensate for features that are missing in one class. Recall that to do so, we pretend to have seen each feature α additional times in each class.

As you increase the Laplacian smoothing parameter α, which of the following will happen to your model?

Laplacian smoothing should only affect features

A) The model's predictions on a test dataset will become more similar to those made without Laplacian smoothing.

B) The model will become equivalent to MAP estimation with a uniform prior distribution.

C) The prior probabilities for each class will eventually approach 0.5.

D) None of the above.

Naive Bayes Formula - $\pi_{y_n} \prod_{n=1}^{N} P(x_n \mid y_n)$ Should only affect these conditional probs.

class prob conditional prob

Gradescope Participation Quiz 02/05

Question 2: As you increase the Laplacian smoothing parameter α, which of the following will happen to your model?

A) The model's predictions on a test dataset will become more similar to those made without Laplacian smoothing.

B) The model will become equivalent to MAP estimation with a uniform prior distribution.

C) The prior probabilities for each class will eventually approach 0.5.

D) None of the above.

Solution: B OR D - Depends on interpretation

Explanation

Answer is contingent on whether or not we consider MAP on only the conditional probabilities or MAP on both the conditional probabilities and the class probabilities.

Linear Regression T/F

Suppose we run linear regression on a dataset with 4 features and 2 data points. The design matrix, weight vector and target vector are as follows (note that we have included a bias term in the model):

$$\boldsymbol{X} = \begin{bmatrix} 1 & -0.5 & 0.5 & 4 & 3 \\ 1 & 7 & 2 & -2 & 3 \end{bmatrix}, \quad \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

Have to determine if this exists

Then the least mean squares solution, i.e., $\boldsymbol{w}^{LMS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, is the unique minimizer of the residual sum of squares for this dataset.

Columns 1 and 5 of X are linearly dependent
⇒ rows 1 and 5 of X^TX will also be linearly dependent
⇒ matrix with L.D. rows cannot be inverted

Linear Regression T/F

Suppose we run linear regression on a dataset with 4 features and 2 data points. The design matrix, weight vector and target vector are as follows (note that we have included a bias term in the model):

$$\boldsymbol{X} = \begin{bmatrix} 1 & -0.5 & 0.5 & 4 & 3 \\ 1 & 7 & 2 & -2 & 3 \end{bmatrix}, \quad \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

Then the least mean squares solution, i.e., $\boldsymbol{w}^{LMS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, is the unique minimizer of the residual sum of squares for this dataset.

Solution: False

Explanation

$\boldsymbol{X}^T\boldsymbol{X}$ is not invertible

Linear Regression MCQ

Let $\boldsymbol{X} \in \mathbb{R}^{(n \times d)}$ be a data matrix with each row representing a datapoint and $\boldsymbol{y} \in \mathbb{R}^n$ be the corresponding vector of labels. Let $\boldsymbol{w}^* = \text{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2$ be the least squares solution and $\boldsymbol{r} = \boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}$ be the *residual error vector*. Note that it might not be possible for $\boldsymbol{w}^*$ to perfectly fit the data, so $\boldsymbol{r} \neq \boldsymbol{0}$ in general. Which of the following statements is true?

A) $\boldsymbol{r}^\top \boldsymbol{w}^* = 0$

→B) $\boldsymbol{r}^\top \boldsymbol{X}\boldsymbol{w}^* = 0$

C) $\boldsymbol{r}^\top \boldsymbol{y} = 0$

D) $\boldsymbol{r}^\top (\boldsymbol{w}^* - \boldsymbol{y}) = 0$

Normal Equation -

$2X^TXw^* - 2X^Ty = 0$

$\Rightarrow X^TXw^* - X^Ty = 0$

$\Rightarrow X^T\underbrace{(Xw^* - y)}_{r} = 0$

$\Rightarrow X^Tr = 0$

$r^TXw^* = \underbrace{(X^Tr)}_{0}{}^T w^*$

So this must be 0

Linear Regression MCQ

Let $\boldsymbol{X} \in \mathbb{R}^{(n \times d)}$ be a data matrix with each row representing a datapoint and $\boldsymbol{y} \in \mathbb{R}^n$ be the corresponding vector of labels. Let $\boldsymbol{w}^* = \text{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2$ be the least squares solution and $\boldsymbol{r} = \boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}$ be the *residual error vector*. Note that it might not be possible for $\boldsymbol{w}^*$ to perfectly fit the data, so $\boldsymbol{r} \neq \boldsymbol{0}$ in general. Which of the following statements is true?

A) $\boldsymbol{r}^\top \boldsymbol{w}^* = 0$

B) $\boldsymbol{r}^\top \boldsymbol{X}\boldsymbol{w}^* = 0$

C) $\boldsymbol{r}^\top \boldsymbol{y} = 0$

D) $\boldsymbol{r}^\top (\boldsymbol{w}^* - \boldsymbol{y}) = 0$

Solution: B

Explanation

Recall the Normal Equation - $2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w}^* - 2\boldsymbol{X}^\top \boldsymbol{y} = 0$. This equivalently means $\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}) = 0 \implies \boldsymbol{X}^\top \boldsymbol{r} = 0$.

Ridge Regression MCQ

For a training dataset $\mathbf{X}_{\text{train}} \in \mathbb{R}^{n \times d}$ and $\mathbf{y}_{\text{train}} \in \mathbb{R}^n$, the optimal ridge regression model is given as

$$\mathbf{w}_\lambda^* = \arg\min_{\mathbf{w}} \left(\|\mathbf{X}_{\text{train}}\mathbf{w} - \mathbf{y}_{\text{train}}\|_2^2 + \lambda \|\mathbf{w}\|_2^2 \right),$$

where λ is the L_2 regularization parameter. Define the train loss and test loss as

$$\mathcal{L}_{\text{train}}(\lambda) = \|\mathbf{X}_{\text{train}}\mathbf{w}_\lambda^* - \mathbf{y}_{\text{train}}\|_2^2,$$

and

$$\mathcal{L}_{\text{test}}(\lambda) = \|\mathbf{X}_{\text{test}}\mathbf{w}_\lambda^* - \mathbf{y}_{\text{test}}\|_2^2,$$

respectively. Which of the following statements are true (more than one can be true)?

A) Adding the L_2 regularization term corresponds to assuming a Gaussian prior on $\mathbf{w}$ True, taken from lecture slides

B) $\mathcal{L}_{\text{train}}(\lambda)$ increases monotonically with respect to λ True, inexactness hurts training loss

C) $\mathcal{L}_{\text{test}}(\lambda)$ decreases monotonically with respect to λ False, too large λ will hurt test loss

D) $\|\mathbf{w}_\lambda^*\|_2^2$ decreases monotonically with respect to λ True large λ will lead to smaller $\|w\|_2$

Ridge Regression MCQ

Which of the following statements are true (more than one can be true)?

A) Adding the L_2 regularization term corresponds to assuming a Gaussian prior on $\mathbf{w}$

B) $\mathcal{L}_{\text{train}}(\lambda)$ increases monotonically with respect to λ

C) $\mathcal{L}_{\text{test}}(\lambda)$ decreases monotonically with respect to λ

D) $\|\mathbf{w}_\lambda^*\|_2^2$ decreases monotonically with respect to λ

Solution: A, B, D

Explanation Ridge regression assumes a Gaussian prior on $\boldsymbol{w}$. The method improves generalization by controlling the magnitude of the entries of $\boldsymbol{w}$ at the expense of the training loss.

MLE/MAP T/F

Suppose that we compute MLE and MAP estimates of the bias of a coin for a dataset of coin tosses, using the Beta(1,2) prior for our MAP estimation. As the dataset size grows larger, the maximum a posteriori probability (MAP) estimator converges to the maximum likelihood estimator (MLE).

$$\hat{\theta}_{MAP} = \frac{1 + n_H - 1}{1 + 2 + n_H + n_T - 1}$$ } Taken from lecture slides

As #trials $\to \infty$, constants don't matter so you recover MLE

Broadly speaking, the more trials you do to analyze the distribution of a p
θ, the less your prior beliefs about θ matter

MLE/MAP T/F

Suppose that we compute MLE and MAP estimates of the bias of a coin for a dataset of coin tosses, using the Beta(1,2) prior for our MAP estimation. As the dataset size grows larger, the maximum a posteriori probability (MAP) estimator converges to the maximum likelihood estimator (MLE).

Solution: True

Explanation

The MAP estimator converges to the MLE as the dataset size grows.

MLE/MAP Descriptive Question

The Geometric distribution is the distribution of the number of independent Bernoulli trials until the first success. In this problem, we will use the Geometric distribution to model an airport baggage claim. Let B be the number of your bag (meaning you watched $B-1$ bags pass before yours arrived). The probability distribution of B is given by

$$P(B=b) = (1-p)^{b-1}p \qquad \text{for } b = 1, 2, \ldots$$

where $0 < p \leq 1$ is the parameter of the distribution.

A) Assume that over n trips, you noted your bag's number as $b_1, b_2, \ldots, b_n$ respectively, and the number of your bag on each of these trips is independent. Derive the log-likelihood function of recording these numbers: $\log p(b_1, \ldots, b_n)$.

B) Derive the maximum likelihood estimate (MLE) of the parameter $\hat{p} = \arg\max_{0<p\leq 1} \log p(b_1, \ldots, b_n)$.

MLE/MAP Descriptive Question

A) Assume that over n trips, you noted your bag's number as $b_1, b_2, \ldots, b_n$ respectively, and the number of your bag on each of these trips is independent. Derive the log-likelihood function of recording these numbers: $\log p(b_1, \ldots, b_n)$.

$\log \mathbb{P}(b_1, \ldots, b_n)$

$= \log[\mathbb{P}(b_1) \ldots \mathbb{P}(b_n)]$ by independence

$= \sum_{i=1}^{n} [\log \mathbb{P}(b_i)]$ by log rules

$= \sum_{i=1}^{n} \log((1-p)^{b_i-1} p)$ by our def. of $\mathbb{P}(b_i)$

$= \sum_{i=1}^{n} [\log(1-p)^{b_i-1} + \log p]$ by log rules

$= n \log p + \sum_{i=1}^{n} \log(1-p)^{b_i-1}$ since $\log p$ does not rely on index i

$= n \log p + \sum_{i=1}^{n} (b_i - 1) \log(1-p)$ by log rules

$= n \log p + \sum_{i=1}^{n} [b_i \log(1-p) - \log(1-p)]$ by distributivity

$= n \log p - n \log(1-p) + \log(1-p) \sum_{i=1}^{n} b_i$ since $\log(1-p)$ does not rely on index i

MLE/MAP Descriptive Question

B) Derive the maximum likelihood estimate (MLE) of the parameter $\hat{p} = \arg\max_{0<p\leq 1} \log p(b_1, \ldots, b_n)$.

To find MLE, take derivative of log likelihood with respect to p and set to 0

$$\frac{\partial \log(b_1, \cdots, b_n)}{\partial p}$$

$$= \frac{\partial[n \log p]}{\partial p} - \frac{\partial[n \log(1-p)]}{\partial p} + \frac{\partial[\log(1-p) \sum_{i=1}^{n} b_i]}{\partial p}$$

$$= \frac{n}{p} + \frac{n}{1-p} - \frac{\sum_{i=1}^{n} b_i}{1-p}$$

$$\Rightarrow \frac{n}{p} + \frac{n}{1-p} - \frac{\sum_{i=1}^{n} b_i}{1-p} = 0$$

$$\Rightarrow \frac{n}{p} + \frac{n}{1-p} = \frac{\sum_{i=1}^{n} b_i}{1-p}$$

$$\Rightarrow \frac{n(1-p)}{p} + n = \sum_{i=1}^{n} b_i$$

$$\Rightarrow \left(\frac{n}{p} - n\right) + n = \sum_{i=1}^{n} b_i$$

$$\Rightarrow \frac{n}{p} = \sum_{i=1}^{n} b_i$$

$$\Rightarrow \hat{p} = \frac{n}{\sum_{i=1}^{n} b_i}$$

Conclusion

Key Takeaways:

Make sure to review: linear algebra concepts, MLE/MAP, linear regression, and bias/variance tradeoff.

Mini Exam will have a combination of T/F, MCQ, and Descriptive questions so don't get stuck on one question for too long.

You are allowed to bring a (one-sided) cheat-sheet for the mini-quiz.

Good luck!