STA314H1F Midterm Review

Chris J. Maddison

October 5, 2025

This Review

- A brief high-level overview of what I think are the key concepts and models.
- We will do two past midterm questions.

High level advice

- Carefully review the homework questions.
- Carefully review the derivations and worked examples in the lectures.
- If, during lecture, I mentioned that a certain derivation is worth doing at home, review that.
- None of the questions will *require* very long derivations, if you can recognize the key insights and intuitions.

Supervised vs. Unsupervised Learning

- **Supervised learning:** Have a collection of training inputs and labels. Goal is to predict label given input.
- **Unsupervised learning:** Have no labeled examples, i.e., only inputs.
- **Regression:** Predicting a scalar-valued label.
- **Classification:** Predicting a discrete-valued label.
- **Decision boundary:** The boundary between regions of input space assigned to different classes by a classifier.

K-Nearest Neighbors (KNN)

- **Idea:** Classify a new input x based on its k nearest neighbors in the training set.
- **Tradeoffs in choosing k:** Overfit vs. Underfit.
- **Pitfalls:** Curse of dimensionality, normalization, computational cost.

Linear Regression

- **Model:** A linear function of the features $y = w^T x$.
- **Loss function:** Squared error loss $\mathcal{L}(y, t) = \frac{1}{2}(y - t)^2$.
- **Average train loss:** Loss averaged over all training examples, i.e. $\hat{\mathcal{R}}[w, \mathcal{D}^{train}]$.
- **Solving:** Direct solution or gradient descent.
- **Gradient Descent Update:** $w \leftarrow w - \alpha \frac{\partial \hat{\mathcal{R}}}{\partial w}$.

Binary Linear Classification

- **Model:**

$$z = w^T x, \quad y = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}$$

- **Geometry:** The model defines a hyperplane decision boundary.
- **Loss Function (0-1 Loss):**

$$\mathcal{L}_{0-1}(y, t) = \mathbb{I}[y \neq t]$$

This loss is non-convex and difficult to optimize. We often use surrogate loss functions.

Logistic Regression (Binary)

- Model: $z = w^T x$
- Loss (Logistic-Cross-Entropy): $\mathcal{L}_{LCE}(z, t) = -t \log(1 + e^{-z}) - (1 - t) \log(1 + e^{z})$.
- To turn a trained logistic regression model into a linear classifier, threshold z using

$$y = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}$$

Decision Trees

- **Model:** Predict by splitting on features in a tree structure.
- **Decision Boundary:** Composed of axis-aligned planes.
- **Fitting strategy:** add splits that maximize information gain.

Information Theory

- **Entropy:** Measures uncertainty in Y.

$$H(Y) = -\sum_{y \in \mathcal{Y}} p(y) \log_2 p(y)$$

- **Conditional Entropy:** Measures uncertainty in Y given X.

$$H(Y|X) = -\sum_{y \in \mathcal{Y}, x \in \mathcal{X}} p(x, y) \log_2 p(y|x)$$

- **Information Gain:** Measures the reduction in entropy in Y after observing X.

$$IG(Y, X) = H(Y) - H(Y|X)$$

- For decision trees, we calculate entropies with respect to the empirical distributions of the labels and splits.

Gradients, Vectorization

- **Gradient:** The column vector of first partial derivatives. For $f : \mathbb{R}^d \to \mathbb{R}$, the gradient is

$$\frac{\partial f}{\partial w} = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \vdots \\ \frac{\partial f}{\partial w_d} \end{bmatrix}$$

- **Vectorization:** re-writing a mathematical expression in terms of vector and matrix operations.

Model Complexity and Generalization

- **Underfitting:** Model is too simplistic to describe the data, high train and test loss.
- **Overfitting:** Model is too complex, fits training data perfectly but fails to generalize to unseen data, high test but low train loss.
- **Hyperparameter:** Can't be included in the training procedure itself; tuned using a validation set.
- **Regularization:** Add a penalty term to the cost function to improve generalization, e.g., L2.

$$\hat{\mathcal{R}}_{reg}[w] = \hat{\mathcal{R}}[w, \mathcal{D}^{train}] + \lambda\phi(w)$$

- **Bias-Variance Decomposition:** decomposed the expected test loss of a trained predictor into three terms, Bayes error, bias, and variance.

Other Things to Know

- Comparisons between different classifiers (KNN, logistic regression, decision trees).
- Contrast the decision boundaries for different classifiers.
- Be adept in the use of dummy variables ($x_0 = 1$) for linear models and the use of feature maps.
- Other topics are fair game: bagging, feature maps, polynomial regression, cross-validation, etc.

2018 Midterm Q7

Question

Consider the classification problem with the following dataset:

x_1	x_2	x_3	t
0	0	0	1
0	1	0	0
0	1	1	1
1	1	1	0

Find a linear classifier with weights w_1, w_2, w_3, and bias w_0 which correctly classifies all examples. No examples should lie on the decision boundary.

(a) Give the set of linear inequalities the weights and bias must satisfy.

(b) Give a setting of the weights and bias that works.

2018 Midterm Q7 - Solution

Part (a): Linear Inequalities

Assuming a dummy variable $x_0 = 1$, for $t = 1$, we need $w^T x + w_0 \geq 0$. For $t = 0$, we need $w^T x + w_0 < 0$. This gives:

$$w_1(0) + w_2(0) + w_3(0) + w_0 \geq 0 \implies w_0 \geq 0$$
$$w_1(0) + w_2(1) + w_3(0) + w_0 < 0 \implies w_2 + w_0 < 0$$
$$w_1(0) + w_2(1) + w_3(1) + w_0 \geq 0 \implies w_2 + w_3 + w_0 \geq 0$$
$$w_1(1) + w_2(1) + w_3(1) + w_0 < 0 \implies w_1 + w_2 + w_3 + w_0 < 0$$

Part (b): Example Weights

Many answers are possible. One corrected solution is:

$$w_1 = -3, \quad w_2 = -2, \quad w_3 = 3, \quad w_0 = 1$$

2018 Midterm Version B Q7

Question

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 Gain Solution

Recall: $IG(Y, X) = H(Y) - H(Y|X)$.

1. **Calculate $H(Y)$:** First, find the marginal probability of Y.
   - $P(Y = 0) = P(X = 0, Y = 0) + P(X = 1, Y = 0) = \frac{1}{8} + \frac{2}{8} = \frac{3}{8}$
   - $P(Y = 1) = P(X = 0, Y = 1) + P(X = 1, Y = 1) = \frac{3}{8} + \frac{2}{8} = \frac{5}{8}$

   Now calculate entropy $H(Y) = -\sum_y P(y) \log_2 P(y)$:

$$H(Y) = -\left(\frac{3}{8}\log_2\frac{3}{8} + \frac{5}{8}\log_2\frac{5}{8}\right)$$

2. **Calculate $H(Y|X)$:** $H(Y|X) = \sum_x P(x)H(Y|X = x)$.
   - $P(X = 0) = \frac{1}{2}$, $P(X = 1) = \frac{1}{2}$
   - $H(Y|X = 0) = -(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4})$
   - $H(Y|X = 1) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2})$

   $H(Y|X) = \frac{1}{2}H(Y|X = 0) + \frac{1}{2}H(Y|X = 1)$

Information Gain Solution (cont.)

3. Combine for Information Gain:

$$
\begin{aligned}
IG(Y, X) &= H(Y) - H(Y|X) \\
&= \left[-\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8}\right] - \\
&\quad \frac{1}{2}\left[-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}\right] - \\
&\quad \frac{1}{2}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right]
\end{aligned}
$$

This is a valid final answer as the question allows for a sum of logarithms.