You'll find our example index cards on the next two pages.

We are not printing them out for you, and you also cannot print them out and bring them to the exam! Remember that the two-sided index card you bring to the exam **must** be handwritten, using no digital tools (no iPads). Feel free to use the material here for inspiration, and/or add your own notes. You may also find these index cards useful while studying.

Step 1: Choose a model.

Constant model: $H(x) = h$

Simple linear regression model: $H(x) = w_0 + w_1 x$

Step 2: Choose a loss function.

Squared loss: $L_{\text{sq}}(y_i, H(x_i)) = (y_i - H(x_i))^2$

Absolute loss: $L_{\text{abs}}(y_i, H(x_i)) = |y_i - H(x_i)|$

Step 3: Minimize average loss (also known as empirical risk) to find optimal model parameters.

If $L(y_i, H(x_i))$ is a loss function, then empirical risk is of the form $R(H) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, H(x_i))$.

Constant model with squared loss: $R_{\text{sq}}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2 \implies h^* = \text{Mean}(y_1, y_2, \ldots, y_n)$

Simple linear regression model with squared loss: $R_{\text{sq}}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2$

$$\implies w_1^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r\frac{\sigma_y}{\sigma_x}$$

$$w_0^* = \bar{y} - w_1^*\bar{x}, \text{ where } r \text{ is the correlation coefficient between } x \text{ and } y.$$

Spans, Projections, and Orthogonality

The span of vectors $\vec{v_1}, \vec{v_2}, \ldots, \vec{v_d} \in \mathbb{R}^n$ is the set of all vectors that can be created using linear combinations of those vectors. A linear combination is of the form $a_1\vec{v_1} + a_2\vec{v_2} + \ldots + a_d\vec{v_d}$, where $a_1, a_2, \ldots, a_d$ are scalars.

Example: If $\vec{v_1} = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}$ and $\vec{v_2} = \begin{bmatrix} -1 \\ 4 \\ 3 \end{bmatrix}$, then $-4\begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix} + 2\begin{bmatrix} -1 \\ 4 \\ 5 \end{bmatrix} = \begin{bmatrix} -10 \\ -4 \\ 14 \end{bmatrix}$ is in span$(\vec{v_1}, \vec{v_2})$.

The span of a single vector $\vec{x} \in \mathbb{R}^n$ is the set of all scalar multiples of $\vec{x}$, i.e. the set of all vectors of the form $w\vec{x}$.

Of all the vectors in span($\vec{x}$), the vector closest to $\vec{y} \in \mathbb{R}^n$ is the vector $w^* \vec{x}$, where:

$$w^* = \frac{\vec{x} \cdot \vec{y}}{\vec{x} \cdot \vec{x}}$$

$w^* \vec{x}$ is called the orthogonal projection of $\vec{y}$ onto span($\vec{x}$). w^* is chosen so that the length of the error vector, $\vec{e} = \vec{y} - w^* \vec{x}$, is minimized. This error vector is orthogonal to $\vec{x}$, i.e. $\vec{x} \cdot \vec{e} = 0$.

Multiple Linear Regression (MLR) and Linear Algebra

The MLR model for a single row of the dataset, i.e. a single data point, is:

$$H(\vec{x}) = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \ldots + w_d x^{(d)} = \vec{w} \cdot \text{Aug}(\vec{x})$$

If $X \in \mathbb{R}^{n \times (d+1)}$ is a design matrix, $\vec{w} \in \mathbb{R}^{d+1}$ is a parameter vector, and $\vec{y} \in \mathbb{R}^n$ is an observation vector, then the predictions for an entire dataset can be written as $\vec{h} = X\vec{w}$, where $\vec{h} \in \mathbb{R}^n$ is a vector containing predictions.

The mean squared error of the MLR model is:

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{y} - X\vec{w}\|$$

The $\vec{w}^*$ that minimizes $R_{\text{sq}}(\vec{w})$ is the $\vec{w}^*$ that satisfies the normal equations, which we found by choosing the error vector $\vec{e} = \vec{y} - X\vec{w}^*$ that is orthogonal to every column in X:

$$X^T(\vec{y} - X\vec{w}^*) = 0 \implies \boxed{X^T X \vec{w}^* = X^T \vec{y}}$$

If $X^T X$ is invertible (equivalently, if the columns of X are all linearly independent), then there is a unique solution to the normal equations: $\vec{w}^* = (X^T X)^{-1} X^T \vec{y}$.