CMPUT 267 Basics of Machine Learning

Prediction and Optimal Predictors
Linear Regression

UNIVERSITY OF ALBERTA

February 29, 2024

Announcements

- Participation and Reading Exercises for MLE and MAP
  - on eClass tonight along with recorded lecture covering MLE and MAP.
  - Exercises will be up for a week.
- Assignment 2a and 2b : deadlines pushed. Check eClass and course website.

Outline

1. Recap
2. Optimal Prediction for Classification - Example
3. Irreducible vs. Reducible Error
4. MLE Formulation for Linear Regression

Recap

- **Supervised learning problem**: Learn a **predictor** $f : \mathcal{X} \to \mathcal{Y}$ from a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$.
  - $\mathcal{X}$ is the set of **observations** and $\mathcal{Y}$ is the set of **targets**.
- **Classification** problems have discrete targets.
- **Regression** problems have continuous targets (order matters).

Recap: Optimal Prediction

- Suppose we know the true joint distribution $p(\mathbf{x}, y)$ and we want to use it to make predictions in a classification problem.
- The **optimal classification predictor** makes the best use of this fucntion.
- As with the optimal estimator, we measure the quality of a predictor $f(\mathbf{x})$ by its **exptected cost** $\mathbb{E}[C]$.
- The optimal predictor **minimizes** $\mathbb{E}[C]$.

$$\mathbb{E}[C] = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} \text{cost}\left(\overbrace{f(\mathbf{x})}^{\hat{y}}, y\right) p(\mathbf{x}, y) d\mathbf{x},$$

where $\text{cost}(\hat{y}, y)$ is the cost of predicting $\hat{y}$ when the true value is y, and $C = \text{cost}\left(f(\mathbf{x}), y\right)$ is a random variable.

Recap: Optimal Classification Prediction

- Bayes risk classifier:

$$f^* = \arg\min_{f \in \mathcal{F}} \int_{\mathcal{X}} p(\mathbf{x})\, \mathbb{E}[C \mid \mathbf{X} = \mathbf{x}] d\mathbf{x}$$

$$f^*(\mathbf{x}) = \arg\min_{f \in \mathcal{F}} \mathbb{E}[C \mid \mathbf{X} = \mathbf{x}] = \arg\min_{\hat{y} \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} \text{cost}(\hat{y}, y) p(y \mid \mathbf{x})$$

- $0 - 1$ cost function:

$$f^* = \arg\max_{\hat{y} \in \mathcal{Y}} p(y \mid \mathbf{x})$$

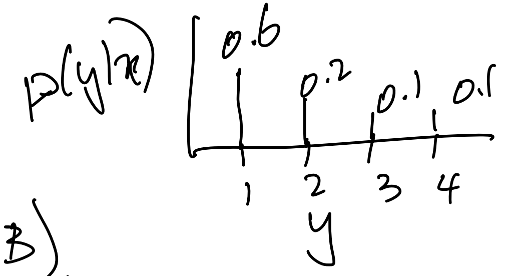
mode of $p(y|x)$

Example

0-1 cost, multiple values for y $\quad y \in \{1, 2, 3, 4\}$

$$cost(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{else} \end{cases}$$

$$f^*(x) = \underset{\hat{y} \in \mathcal{Y}}{\arg\min}\ E[Cost \mid X = x] = \underset{\hat{y} \in \mathcal{Y}}{\arg\max}\ \underline{p(y \mid x)}$$



$$E[Cost \mid X = x]$$
$$= 0.6\, Cost(\hat{y}, 1) + 0.2\, Cost(\hat{y}, 2) + 0.1\, cost(\hat{y}, 3) + 0.1\, cost(\hat{y}, 4)$$

if $\hat{y} = 1$: $E[Cost \mid X = x] = 0.6 \cdot 0 + 0.2 \cdot 1 + 0.1 \cdot 1 + 0.1 \cdot 1 = 0.4$ $\quad \hat{y} = 1$: lowest cost

if $\hat{y} = 2$ $\quad = 0.6 \cdot 1 + 0.2 \cdot 0 + 0.1 \cdot 1 + 0.1 \cdot 1 = 0.8$ . . .

Recap: Optimal Regression Prediction

- Most common cost functions:
  1. squared error: $\text{cost}(\hat{y}, y) = (\hat{y} - y)^2$
  2. absolute error: $\text{cost}(\hat{y}, y) = |\hat{y} - y|$
- Squared error function penalizes large error values more than absolute error function.
- Optimal prediction for squared error:

$$f^*(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$$

For absolute error $f^*(x) = \text{median}(Y|x)$

Example: Classification

- A medical example where "$0-1$" cost, but with cost differing by type of wrong answer.
- $y=-1$: no disease; $y=1$: disease.
- if disease predicted, further tests; no tests if no disease predicted.
- false positive: leads to unnecessary test.
- false negative: leads to untreated disease (and law suit later).

Y - true

$\hat{Y}$ prediction	-1 (no disease)	1 (disease)
-1 (no disease)	0 TN	999 FN
$+1$ (disease)	1 FP	1 TP

0

Example

- $y \in \{-1, 1\}$
- Given $\mathbf{x}$, let:

$$p(y = 1 \mid \mathbf{x}) = p_1 \quad p(y = -1 \mid \mathbf{x}) = p_0 \quad (p_0 = 1 - p_1)$$

$$f^*(\mathbf{x}) = \arg\min_{f \in \mathcal{F}} \mathbb{E}[C \mid \mathbf{X} = \mathbf{x}]$$

$$f^*(x) = \arg\min_{\hat{y} \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} \text{cost}(\hat{y}, y) p(y \mid x)$$

$$f^* = \underset{\hat{y} \in \mathcal{Y}}{\operatorname{argmin}} \sum_{y \in \mathcal{Y}} \operatorname{cost}(\hat{y}, y)\, p(y|x) = \underset{\hat{y} \in \mathcal{Y}}{\operatorname{argmin}} \left[\operatorname{cost}(\hat{y}, y=1)\, p(y=1|x) + \operatorname{cost}(\hat{y}, y=-1)\, p(y=-1|x) \right]$$

$\hat{y} = +1$ or -1

$$= \underset{\hat{y} \in \mathcal{Y}}{\operatorname{argmin}} \underbrace{\operatorname{cost}(\hat{y}, y=1)\, p_1 + \operatorname{cost}(\hat{y}, y=-1)\, p_0} \rightarrow \ell(\hat{y})$$

if $\hat{y} = -1$ $\quad \ell(\hat{y} = -1) = 999 p_1 + 0 \cdot p_0 = 999 p_1$

if $\hat{y} = +1$ $\quad \ell(\hat{y} = 1) = 1 \cdot p_1 + 1 \cdot p_0 = 1$

$$\boxed{f^* = \underset{\hat{y} \in \mathcal{Y}}{\operatorname{argmin}}\ \ell(\hat{y})}$$

If $999 p_1 < 1$ $\quad f^* : \hat{y} = -1$ $\quad \left[\text{if } p_1 < \frac{1}{999}\right]$

Else $\quad f^* : \hat{y} = +1$

$p_1 = P(y=1|x)$

Error in prediction

Regression

- Optimal prediction doesn't mean error =0.
- What is the quality of our predictor? It may be optimal or suboptimal. Let's look at the expected squared error.
- First let's consider the optimal predictor, $f^*(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$.

$$\begin{aligned}\mathbb{E}[C] &= \int_{\mathcal{X}} p(\mathbf{x}) \boxed{\int_{\mathcal{Y}}} (f^*(\mathbf{x}) - y)^2 p(y \mid \mathbf{X} = \mathbf{x}) dy dx \\ &= \int_{\mathcal{X}} p(\mathbf{x}) \int_{\mathcal{Y}} (\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}] - y)^2 p(y \mid \mathbf{X} = \mathbf{x}) dy dx\end{aligned}$$

Error in prediction

- Optimal prediction doesn't mean error =0.
- What is the quality of our predictor? It may be optimal or suboptimal. Let's look at the expected squared error.
- First let's consider the optimal predictor, $f^*(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$.

$$\begin{aligned}
\mathbb{E}[C] &= \int_{\mathcal{X}} p(\mathbf{x}) \int_{\mathcal{Y}} (f^*(\mathbf{x}) - y)^2 p(y \mid \mathbf{X} = \mathbf{x}) dy dx \\
&= \int_{\mathcal{X}} p(\mathbf{x}) \int_{\mathcal{Y}} (\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}] - y)^2 p(y \mid \mathbf{X} = \mathbf{x}) dy dx \\
&= \int_{\mathcal{X}} p(\mathbf{x}) \operatorname{Var}[Y \mid \mathbf{X} = \mathbf{x}]
\end{aligned}$$

- This is **irreducible error**.

Error for any predictor

$(a+b)^2$

Now let's consider the expected square error for a suboptimal predictor, $f(\mathbf{x})$.

$$\mathbb{E}[C \mid X] = \mathbb{E}\left[(f(\mathbf{x}) - Y)^2 \mid X = \mathbf{x}\right] = \mathbb{E}\left[(\underbrace{f(\mathbf{x}) - \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]}_{a} + \underbrace{\mathbb{E}[Y \mid X = \mathbf{x}] - Y}_{b})^2 \mid \mathbf{X} = \mathbf{x}\right]$$

$$\begin{aligned} &= \mathbb{E}\Big[\underbrace{(f(\mathbf{x}) - \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}])^2}_{a^2} + \underbrace{2\,(f(\mathbf{x}) - \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}])(\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}] - Y)}_{2ab} \\ &\qquad + \underbrace{(\mathbb{E}[Y \mid X = \mathbf{x}] - Y)^2}_{b^2} \mid \mathbf{X} = \mathbf{x}\Big] \end{aligned}$$

Error for any predictor

Now let's consider the expected square error for a suboptimal predictor, $f(\mathbf{x})$.

$$\mathbb{E}[C \mid X] = \mathbb{E}\left[(f(\mathbf{x}) - Y)^2 \mid X = \mathbf{x}\right] = \mathbb{E}\left[(f(\mathbf{x}) - \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}] + \mathbb{E}[Y \mid X = \mathbf{x}] - Y)^2 \mid \mathbf{X} = \mathbf{x}\right]$$

$$= \mathbb{E}\Big[(f(\mathbf{x}) - \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}])^2 + 2\boxed{(f(\mathbf{x}) - \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}])(\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}] - Y)}$$

$$+ (\mathbb{E}[Y \mid X = \mathbf{x}] - Y)^2 \mid \mathbf{X} = \mathbf{x}\Big]$$

Error (cont'd), middle term

ab

$$\mathbb{E}[(f(\mathbf{x}) - \mathbb{E}[Y \mid X = \mathbf{x}])(\mathbb{E}[Y \mid X = \mathbf{x}] - Y) \mid \mathbf{X} = \mathbf{x}]$$

$$= (f(\mathbf{x}) - \mathbb{E}[Y \mid X = \mathbf{x}])\, \mathbb{E}\left[(\mathbb{E}[Y \mid X = \mathbf{x}] - Y) \,\middle|\, X = \mathbf{x}\right]$$

$$= (f(\mathbf{x}) - \mathbb{E}[Y \mid X = \mathbf{x}])\, (\mathbb{E}[Y \mid X = \mathbf{x}] - \mathbb{E}[Y|X = \mathbf{x}])$$

$$= (f(\mathbf{x}) - \mathbb{E}[Y \mid X = \mathbf{x}])\, 0$$

$$= 0$$

Error for any predictor

$$E[C] = \int_{x \in \mathcal{X}} p(x)\, E[C \mid X = x]\, dx$$

$$\begin{aligned}
\mathbb{E}[C] &= \mathbb{E}[\mathbb{E}[C \mid \mathbf{X} = \mathbf{x}]] \\
&= \mathbb{E}\left[\mathbb{E}\left[\overset{a^2}{(f(\mathbf{x}) - \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}])^2} + \overset{b^2}{(\mathbb{E}[Y \mid X = \mathbf{x}] - Y)^2} \mid \mathbf{X} = \mathbf{x}\right]\right] \\
&= \mathbb{E}\left[(f(\mathbf{X}) - \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}])^2\right] + \mathbb{E}\left[(\mathbb{E}[Y \mid X = \mathbf{x}] - Y)^2 \mid \mathbf{X} = \mathbf{x}\right] \\
&= \mathbb{E}\left[(f(\mathbf{X}) - f^*(\mathbf{X}))^2\right] \quad + \quad \mathbb{E}\left[(f^*(\mathbf{X}) - Y)^2\right]
\end{aligned}$$

Error for any predictor

$$
\begin{aligned}
\mathbb{E}[C] &= \mathbb{E}[\mathbb{E}[C \mid \mathbf{X} = \mathbf{x}]] \\
&= \mathbb{E}\left[\mathbb{E}\left[(f(\mathbf{x}) - \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}])^2 + (\mathbb{E}[Y \mid X = \mathbf{x}] - Y)^2 \mid \mathbf{X} = \mathbf{x}\right]\right] \\
&= \mathbb{E}\left[(f(\mathbf{X}) - \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}])^2\right] + \mathbb{E}\left[(\mathbb{E}[Y \mid X = \mathbf{x}] - Y)^2 \mid \mathbf{X} = \mathbf{x}\right] \\
&= \boxed{\mathbb{E}\left[(f(\mathbf{X}) - f^*(\mathbf{X}))^2\right]} + \boxed{\mathbb{E}\left[(f^*(\mathbf{X}) - Y)^2\right]}
\end{aligned}
$$

reducible error irreducible error

How to reduce the reducible error?

- We want to make the error between the f that we learn and the optimal f^* smaller.
- Let's assumed the hypothesis space we're looking in, $\mathcal{F}$, is the space of linear functions.
- **Sources of reducible error**
  1. **Limited hypothesis space**. We assumed linear functions, but maybe f^* is non linear.
  2. **Insufficient optimization**. We might have used gradient descent, but did't fully optimize f - stopped too early?
  3. **Limited data**. Not enough samples to identify a good f.

How to reduce the reducible error?

1. **Limited hypothesis space**. We assumed linear functions, but maybe f^* is non linear.
   - ▷ Solution: make the hypothesis space bigger (e.g. polynomials?)
2. **Insufficient optimization**. We might have used gradient descent, but did't fully optimize f - stopped too early?
   - ▷ Solution: set step size and number of epochs more carefully to ensure you're at a stationary point.
3. **Limited data**. Not enough samples to identify a good f.
   - ▷ Solution: gather more data.

How to reduce the irreducible error?

$\int p(x) \, Var[Y | X=x]$

- It's **irreducible**...
- It's the variance of Y given X: $\mathrm{Var}(Y \mid \mathbf{X} = \mathbf{x})$.
- Improving the learned function **cannot** change the inherent variance in Y.
- BUT: what is the source of variance in Y given x?

How to reduce the irreducible error?

- It's **irreducible**...
- It's the variance of Y given X: $\mathrm{Var}(Y \mid \mathbf{X} = \mathbf{x})$.
- Improving the learned function **cannot** change the inherent variance in Y.
- BUT: what is the source of variance in Y given x?
  - partial observability
  - stochasticity in the system

Linear Predictors

Linear Regression

Setting : $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ $\quad x_i \in \mathbb{R}^d \quad y_i \in \mathbb{R}$

$$f(x_i) = \underline{w_0} + w_1 x_{i,1} + w_2 x_{i,2} + \cdots + w_d x_{i,d}$$

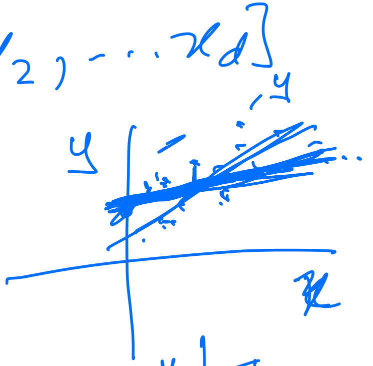
$$f(x_i) = \sum_{j=0}^{d} w_j x_{i,j} \qquad (x_{i,0} = 1)$$

redefine x : $[1, x_1, x_2, \ldots x_d]$

$p(y \mid x)$ $\quad d = 1 \quad w_0 = 0$

$$p(y|x) = N(\underline{x w}, \sigma^2)$$

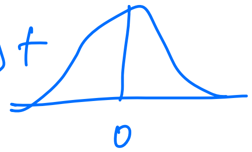
$$\underline{p(y|x)} \leftarrow N(\mu = f^*(x), \sigma^2)$$



$$\left\{ \underbrace{y_i = \sum_{j=0}^{d} w_j x_{i,j}}_{w^T x} + \underline{\varepsilon_i} \right. \qquad \varepsilon_i \sim N(0, \sigma^2)$$

$x^T w - x^T w$

$E[y \mid x = x] +$



$\rightarrow$

Hypothesis space:

$$\mathcal{F} = \{ p(\cdot \mid x) = \mathcal{N}(w^T x, \sigma^2) \mid w \in \mathbb{R}^d \}$$

Goal: find parameters w_j, $j = 0, \ldots, d$

$\vec{w} \in \mathbb{R}^{d+1}$

that is optimal in $\mathcal{F}$

MLE : $\underset{w \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i=1}^{n} -\ln p(y_i \mid x_i, \vec{w})$

MLE for Linear regression

$$\underset{w \in \mathbb{R}^{d+1}}{\arg\max}\; p(D|w) = \underset{w \in \mathbb{R}^{d+1}}{\arg\min} -\ln\left(p(D|w)\right)$$

$$\boxed{c(w) = \frac{1}{n}\sum_{i=1}^{n} c_i(w)}$$

$$= \underset{w \in \mathbb{R}^{d+1}}{\arg\min}\; \frac{1}{n}\left(\sum_{i=1}^{n} -\ln p(y_i | \vec{x}_i, \vec{w})\right) \quad \mathcal{N}(x^T w, \sigma^2)$$

$$= \underset{w \in \mathbb{R}^{d+1}}{\arg\min} -\frac{1}{n}\sum_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x^T w)^2}{2\sigma^2}\right)$$

$$= \underset{w \in \mathbb{R}^{d+1}}{\arg\min}\; \frac{1}{n}\sum_{i=1}^{n} \frac{(x^T w - y_i)^2}{2} = \boxed{\underset{w \in \mathbb{R}^{d+1}}{\arg\min}\; \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(x^T w - y_i)^2}$$

Ordinary Least Squares

Find w_{MLE} $\quad \nabla c(w) = 0$

$$\begin{bmatrix} \frac{\partial}{\partial w_0} c(w) \\ \vdots \\ \frac{\partial}{\partial w_d} c(w) \end{bmatrix}$$

$$\frac{\partial}{\partial w_j} c(w) = \frac{1}{n} \sum \underbrace{\frac{\partial c_i(w)}{\partial w_j}}$$

$$\hookrightarrow \frac{\partial c_i(w)}{\partial w_j} = \underbrace{\frac{\partial \frac{1}{2}(x^T w - y_i)^2}{\partial(x^T w)}} \; \underbrace{\frac{\partial (x^T w)}{\partial w_j}}$$

$$\frac{1}{n} \sum_{i=1}^{n} (x^T w - y_i)\, x_i = 0$$

$$\hookrightarrow \frac{\partial (x^T w)}{\partial w_j} = \frac{\partial}{\partial w_j} \sum_{k=0}^{d} x_{ik} w_k$$

$$= \sum_{k=0}^{d} x_{ik} \frac{\partial w_k}{\partial w_j} = x_{ij}$$

$$\frac{1}{n}\sum(x_i^T w - y_i)\, x_{i0} = 0$$

$$\frac{1}{n}\sum(x_i^T w - y_i)\, x_{i1} = 0 \quad \longrightarrow \quad Aw = b$$

$$\vdots$$

$$\frac{1}{n}\sum(x_i^T w - y_i)\, x_{id} = 0$$

$$A = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \qquad b = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

$O(nd^2)$ matrix inversion $O(d^3)$

$O(nd^2 + d^3)$

Numerical Solution $\vec{w}_t$

SGD. 1. Initialize $\vec{w}_0$ $w_{0,j} = N(0,1)$

2. Sample minibatch (shuffle) b: batch size

3. $\vec{w}_{t+1} = \vec{w}_t - \eta_t \underbrace{\frac{1}{b} \sum_{i=1}^{b} (x_i^T \vec{w}_t - y_i) x_i}_{g_t}$

η_t (vector)

$\bar{g}_{t,j} = \bar{g}_{t-1,j} + g_{t,j}^2$

$\eta_{t,j} = \frac{1}{10^{-7} + \sqrt{\bar{g}_{t,j}}}$

$\bar{g}_0 = 0$

numerical issues at initialization

t

$10^{-4} t$

w_{MLE}

$O(kbd)$
(for k iteration of SGD)

$$w_{t+1} = w_t - \eta_t \, c'(w_t)$$

$$\frac{1}{n}\sum_{i=1}^{n} c_i'(w_t)$$

$b \ll n$ $\quad \frac{1}{b}\sum_{i=1}^{b} c_i'(w)$