Optimal Prediction cont. & Linear Regression

CMPUT 296: Basics of Machine Learning

Textbook §6.2, §7.1

Logistics

1. **"In-class" quiz Thursday Oct 8** (next week!)
    - Covers all material through section 7.1
    - Tuesday class will be a review
    - Quiz will be on eClass during a 24 hour period
    - Random spot checks scheduled starting the following week
2. **Thought questions #2** also due **October 8**
    - TQ#1 will be marked by the end of this week

Recap: Supervised Learning

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

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 function.

As with the optimal estimator, we measure the quality of a predictor $f(\mathbf{x})$ by its **expected cost** $\mathbb{E}[C]$. The optimal predictor **minimizes** $\mathbb{E}[C]$.

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

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

Outline

1. Recap & Logistics
2. Optimal Prediction for Regression
3. Irreducible and Reducible Error
4. MLE Formulation for Linear Regression

Cost Functions: Regression

- Two most common cost functions for regression:
  1. **Squared error:** $\mathrm{cost}(\hat{y}, y) = (\hat{y} - y)^2$
  2. **Absolute error:** $\mathrm{cost}(\hat{y}, y) = \left|\hat{y} - y\right|$
- Squared error penalizes **large errors** more heavily than absolute error
- Other possibilities that depend on the size of the target
  - E.g., **percentage error:** $\mathrm{cost}(\hat{y}, y) = \dfrac{\left|\hat{y} - y\right|}{|y|}$

Deriving Optimal Regressor for Squared Error

$$\mathbb{E}[C] = \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathrm{cost}\big(f(\mathbf{x}), y\big)\, p(\mathbf{x}, y)\, dy\, d\mathbf{x}$$

$$= \int_{\mathcal{X}} \int_{\mathcal{Y}} \big(f(\mathbf{x}) - y\big)^2 p(\mathbf{x}, y)\, dy\, d\mathbf{x}$$

$$= \int_{\mathcal{X}} p(\mathbf{x}) \underbrace{\boxed{\int_{\mathcal{Y}} \big(f(\mathbf{x}) - y\big)^2 p(y \mid \mathbf{x})\, dy}}_{\mathbb{E}[C \mid X = \mathbf{x}]} d\mathbf{x}$$

$$= \int_{\mathcal{X}} p(\mathbf{x}) \mathbb{E}[C \mid X = \mathbf{x}]\, d\mathbf{x}$$

- Once again, we can directly optimize $\mathbb{E}[C \mid X = \mathbf{x}]$:

$$f^*(\mathbf{x}) = \arg\min_{\hat{y} \in \mathcal{Y}} g(\hat{y})$$

where

$$g(\hat{y}) = \int_{\mathcal{Y}} \big(\hat{y} - y\big)^2 p(y \mid \mathbf{x})\, dy$$

Deriving Optimal Regressor for Squared Error, cont.

$$g(\hat{y}) = \int_{\mathcal{Y}} (\hat{y} - y)^2 \, p(y \mid \mathbf{x}) \, dy$$

$$\frac{\partial g(\hat{y})}{\partial \hat{y}} = 2 \int_{\mathcal{Y}} (\hat{y} - y) \, p(y \mid \mathbf{x}) \, dy = 0$$

$$\iff \int_{\mathcal{Y}} \hat{y} p(y \mid \mathbf{x}) \, dy = \int_{\mathcal{Y}} y p(y \mid \mathbf{x}) \, dy$$

$$\iff \hat{y} \boxed{\underbrace{\int_{\mathcal{Y}} p(y \mid \mathbf{x}) \, dy}_{= 1}} = \int_{\mathcal{Y}} y p(y \mid \mathbf{x}) \, dy$$

$$\iff \hat{y} = \boxed{\int_{\mathcal{Y}} y p(y \mid \mathbf{x}) \, dy} = \mathbb{E}[Y \mid X = \mathbf{x}]$$

So,

$$f^*(\mathbf{x}) = \arg\min_{\hat{y} \in \mathcal{Y}} g(\hat{y})$$

$$= \mathbb{E}[Y \mid X = \mathbf{x}] \quad \blacksquare$$

Generative Models

- The optimal prediction approach depends on (an estimate of) $p(y \mid \mathbf{x})$
- Two approaches to learning $p(y \mid \mathbf{x})$:
  1. **Discriminative:** Learn $p(y \mid \mathbf{x})$ directly
  2. **Generative:** Learn $p(\mathbf{x} \mid y)$ and $p(y)$,
     and exploit $p(y \mid \mathbf{x}) \propto p(\mathbf{x} \mid y)p(y)$
- **Question:** What are the relative advantages of these two approaches?

Irreducible Error

What is our **expected squared error** when we use the **optimal** predictor?

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

$$
\begin{aligned}
\mathbb{E}[C] &= \int_{\mathscr{X}} p(\mathbf{x}) \int_{\mathscr{Y}} \left(f^*(\mathbf{x}) - y\right)^2 p(y \mid X = \mathbf{x})\, dy\, d\mathbf{x} \\
&= \int_{\mathscr{X}} p(\mathbf{x}) \boxed{\int_{\mathscr{Y}} \left(\mathbb{E}[Y \mid X = \mathbf{x}] - y\right)^2 p(y \mid X = \mathbf{x})\, dy}\, d\mathbf{x} \\
&= \int_{\mathscr{X}} p(\mathbf{x}) \mathrm{Var}[Y \mid X = \mathbf{x}]\, d\mathbf{x}
\end{aligned}
$$

Reducible Error

What is our **expected squared error** when we use a **suboptimal** predictor?

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

$$= \mathbb{E}\left[\left(f(\mathbf{x}) - \mathbb{E}[Y \mid X = \mathbf{x}]\right)^2 + 2\underbrace{\boxed{\left(f(\mathbf{x}) - \mathbb{E}[Y \mid X = \mathbf{x}]\right)\left(\mathbb{E}[Y \mid X = \mathbf{x}] - Y\right)}}_{=0}\right.$$

$$\left. + \left(\mathbb{E}[Y \mid X = \mathbf{x}] - Y\right)^2 \,\middle|\, X = \mathbf{x}\right]$$

We'll take expectation again at the end to get to $\mathbb{E}[C] = \mathbb{E}[\mathbb{E}[C \mid X]]$

Reducible Error: Middle Term is 0

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

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

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

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

$$= 0$$

Reducible Error

What is our **expected squared error** when we use a **suboptimal** predictor?

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

$$= \mathbb{E}\left[\left(f(\mathbf{x}) - \mathbb{E}[Y \mid X = \mathbf{x}]\right)^2 + 2\underbrace{\boxed{\left(f(\mathbf{x}) - \mathbb{E}[Y \mid X = \mathbf{x}]\right)\left(\mathbb{E}[Y \mid X = \mathbf{x}] - Y\right)}}_{= 0}\right.$$

$$\left. + \left(\mathbb{E}[Y \mid X = \mathbf{x}] - Y\right)^2 \,\middle|\, X = \mathbf{x}\right]$$

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

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

$$\mathbb{E}[C] = \underbrace{\boxed{\mathbb{E}\left[\left(f(X) - f^*(X)\right)^2\right]}}_{\text{Reducible error}} + \underbrace{\boxed{\mathbb{E}\left[\left(f^*(X) - Y\right)^2\right]}}_{\text{Irreducible error}}$$

Summary

- **Supervised learning problem:** Learn a **predictor** $f : \mathscr{X} \to \mathscr{Y}$ from a dataset $\mathscr{D} = \left\{ (\mathbf{x}_i, y_i) \right\}_{i=1}^{n}$
  - $\mathscr{X}$ is the set of **observations**, and $\mathscr{Y}$ is the set of **targets**
- **Classification** problems have discrete targets
- **Regression** problems have continuous targets
- Predictor performance is measured by the **expected** $\text{cost}(\hat{y}, y)$ of predicting $\hat{y}$ when the true value is y
- An **optimal predictor** for a given distribution **minimizes** the expected cost
- Even an optimal predictor has some **irreducible error**.
  Suboptimal predictors have additional, **reducible error**

Linear Predictors

A **linear predictor** is a function of the form

$$f(\mathbf{x}) = w_0 + w_1 x_1 + \ldots + w_d x_d = \sum_{j=\boxed{0}}^{d} w_j x_j = \mathbf{w}^T \mathbf{x}$$

Intercept feature $x_0 = 1$

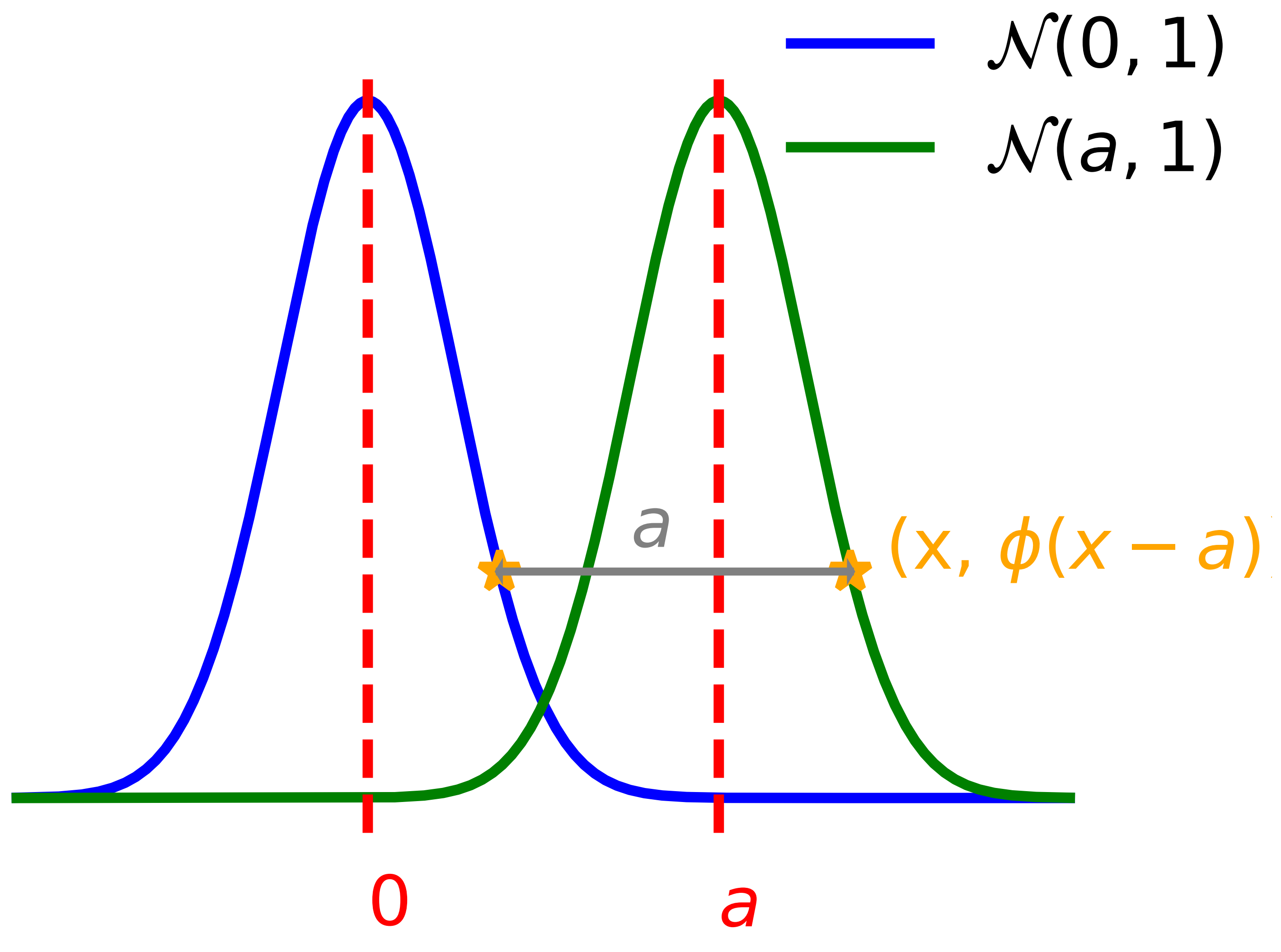
- Predict a **linear combination** of **weights** $\mathbf{w}$ and **features** $\mathbf{x}$
- **Linear regression:** finding the best parameters $\mathbf{w} \in \mathbb{R}^{d+1}$
- **Question:** What criterion should we use to pick $\mathbf{w}$?

Gaussian Error Model

Suppose that our dataset $\mathscr{D} = \left\{(\mathbf{x}_i, y_i)\right\}_{i=1}^{n}$ was drawn as follows:

1. $\mathbf{x}_i \overset{i.i.d.}{\sim} p(\mathbf{x})$
2. $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$
3. $y_i = \sum_{j=0}^{d} \omega_j x_{ij} + \epsilon_i$

Question: what is the distribution of Y?



$$Y = \sum_{j=0}^{n} \omega_j x_{ij} + \epsilon$$

$$Y \sim \mathcal{N}(\omega^T \mathbf{x}, \sigma^2)$$

Linear Regression as Model Estimation

- We now have a **parametric family of conditional models** to select from:

$$\mathscr{F} = \left\{p(\,\cdot \mid \mathbf{x}) = \mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma^2) \mid \mathbf{w} \in \mathbb{R}^{d+1}\right\}.$$

  - (Equivalently, need to select a parameter vector $\mathbf{w} \in \mathbb{R}^{d+1}$ that identifies a conditional model in the family)

- Once we have selected a model, we can use it to make predictions

- **Question:** *How* should we use an estimated model $p(y \mid \mathbf{x}, \mathbf{w})$ for prediction?

MLE for Linear Regression

$$\mathbf{w}_{\text{MLE}} = \arg\max_{\mathbf{w}\in\mathbb{R}^{d+1}} p(\mathcal{D} \mid \mathbf{w})$$

$$= \arg\max_{\mathbf{w}\in\mathbb{R}^{d+1}} \prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$$= \arg\max_{\mathbf{w}\in\mathbb{R}^{d+1}} \ln\left(\prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i, \mathbf{w})\right)$$

$$= \arg\max_{\mathbf{w}\in\mathbb{R}^{d+1}} \sum_{i=1}^{n} \ln p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} -\sum_{i=1}^{n} \ln p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} -\sum_{i=1}^{n} \ln \left[\underbrace{\boxed{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i - \mathbf{w}^T\mathbf{x})^2}{2\sigma^2}\right)}}_{\mathcal{N}(\mathbf{w}^T\mathbf{x}_i, \sigma^2)(y_i)}\right]$$

$\exp(a) = e^a$

MLE for Linear Regression cont.

$$\mathbf{w}_{\text{MLE}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} -\sum_{i=1}^{n} \ln\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-\mathbf{w}^T\mathbf{x})^2}{2\sigma^2}\right)\right]$$

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

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \boxed{\sum_{i=1}^{n}\ln\sqrt{2\pi\sigma^2}} + \sum_{i=1}^{n}\frac{(y_i-\mathbf{w}^T\mathbf{x})^2}{2\sigma^2}$$

Constant w.r.t. $\mathbf{w}$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \sum_{i=1}^{n}\frac{(y_i-\mathbf{w}^T\mathbf{x})^2}{2\sigma^2}$$

Constant w.r.t. $\mathbf{w}$

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

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \sum_{i=1}^{n}(y_i-\mathbf{w}^T\mathbf{x})^2 \quad \blacksquare$$

Prediction with MLE Model

We have an **estimated model** of the process: $Y \sim \mathcal{N}\left(\mathbf{w}_{\text{MLE}}^T \mathbf{x}, \sigma^2\right)$,

where $\mathbf{w}_{\text{MLE}} = \arg\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x})^2$

Question: How should we use this estimated model for prediction?

- Use the **optimal regression predictor** assuming this is the correct model:

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

Question: What is the expected value of Y conditional on $X = \mathbf{x}$?

Ordinary Least Squares

$$\mathbf{w}_{\text{MLE}} = \arg \min_{\mathbf{w} \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x})^2$$

We just minimize the sum of squared errors on our dataset $\mathscr{D}$!

Question: What are the advantages of doing the MLE derivation rather than just directly minimizing error on the dataset?

1. It makes the **assumptions** behind the process clear:
   - underlying linear relationship between y_i and $\mathbf{x}_i$
   - i.i.d. errors in y_i
   - no errors in $\mathbf{x}_i$
   - **noise** (error term) ϵ_i is a zero-mean Gaussian
   - noise ϵ_i is independent of the features $\mathbf{x}_i$

2. It is a **general** approach:
   - A good objective for other distributions of $p(y \mid \mathbf{x})$ is not as obvious
   - But the MLE approach will work for any distribution

Summary

A **linear predictor** has the form $f(\mathbf{x}) = w_0 + w_1 x_1 + \ldots + w_d x_d = \sum_{j=0}^{d} w_j x_j = \mathbf{w}^T \mathbf{x}$

Traditional approach: Find the linear predictor that minimizes squared error on the dataset (aka **Ordinary Least Squares**)

Probabilistic approach:

1. Assume **i.i.d. Gaussian noise**: $Y \sim \mathcal{N}(\omega^T \mathbf{x}, \sigma^2)$
2. Use MLE to estimate model from resulting **parametric family** $\mathcal{F} = \left\{ p(\cdot \mid \mathbf{x}) = \mathcal{N}(\mathbf{w}^T \mathbf{x}, \sigma^2) \mid \mathbf{w} \in \mathbb{R}^{d+1} \right\}$
3. Use the **optimal predictor** for the estimated model $\mathbf{w}^*$: $f^*(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}] = \mathbf{w}^T \mathbf{x}$