1. Simple prediction problems

Prediction problem #1

A coin is tossed.

Our task: predict the outcome (either “heads” or “tails”).

How should we predict?

1. Physical model
   
2. Statistical model
   Assume outcome is random:
   “heads” with probability $p$, “tails” with probability $1 - p$. 

Prediction strategy for problem #1

Suppose we know $p$. How should we predict?

- If $p > 1/2$, then predict “heads”.
- If $p < 1/2$, then predict “tails”.
- If $p = 1/2$, doesn’t matter. But, for concreteness, predict “tails”.

Using this strategy, what is the probability that you predict incorrectly? Is it possible to do any better?

If we encode “heads” = 1 and “tails” = 0, we say outcome is a Bernoulli random variable $Y \sim \text{Bern}(p)$. 

Prediction problem #2

A ball is dropped in a Galton board.\(^1\)

Our task: predict the (horizontal) position of the ball when it lands.

(Assume we have agreed on a coordinate system.)

Quality of prediction \(\hat{y}\) assessed by loss function.

We’ll use squared loss \((\hat{y} - y)^2\).

\(^1\)You can see one at the New York Hall of Science!

Model for problem #2

Statistical model: outcome is \(Y \sim N(\mu, \sigma^2)\), a normal distribution.

Parameters \(\mu \in \mathbb{R}, \sigma^2 > 0\).

Probability density function (pdf) for \(Y\) is

\[
\phi_{\mu, \sigma^2}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right), \quad y \in \mathbb{R}.
\]

Central moments of \(Y\): \(\mathbb{E}(Y) = \mu, \text{var}(Y) = \sigma^2, \ldots\)

Prediction strategy for problem #2

Suppose we know \((\mu, \sigma^2)\). How should we predict?

If we predict \(\hat{y}\), what is the expected loss (a.k.a. risk) \(\mathcal{R}(\hat{y}) := \mathbb{E}((\hat{y} - Y)^2)]\)?

For any \(\hat{y} \in \mathbb{R},\)

\[
\mathcal{R}(\hat{y}) = \mathbb{E}((\hat{y} - Y)^2]
= \mathbb{E}((\hat{y} - \mu + \mu - Y)^2]
= (\hat{y} - \mu)^2 + 2(\hat{y} - \mu) \mathbb{E}(\mu - Y)] + \mathbb{E}(\mu - Y)^2]
= (\hat{y} - \mu)^2 + \mathbb{E}(\mu - Y)^2]
= (\hat{y} - \mu)^2 + \text{var}(Y).
\]

So what is the best prediction \(\hat{y}\)?

Actually, this does not require \(Y\) to be normally distributed; it is a consequence of using squared loss.

2. From data to predictions
What if we don’t know model parameters?

Often, we don’t know model parameters (e.g., \( p, (\mu, \sigma^2) \)), ... but we see related observations (data) before we need to make prediction (e.g., previous balls dropped in Galton board).

**Plug-in principle:**
1. Estimate unknowns based on data.
2. Plug these estimates into formula.

But how is data related to outcome?

**IID model:** Observations & outcome are independent & identically distributed (iid) random variables.

---

**Maximum likelihood estimation**

**Parametric statistical model:**
\( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \), a collection of probability distributions for observed data.
- \( \Theta \): parameter space.
- \( \theta \in \Theta \): a particular parameter (or parameter vector).
- \( P_\theta \): a particular probability distribution for observed data.

**Likelihood of** \( \theta \in \Theta \) **given observed data** \( x \):
For continuous \( X \sim P_\theta \) with probability density function \( f_\theta \),
\[
\mathcal{L}(\theta) := f_\theta(x).
\]
(For discrete \( X \), use probability mass function.)

**Maximum likelihood estimator (MLE):**
Let \( \hat{\theta} \) be the \( \theta \in \Theta \) of highest likelihood given observed data.

---

**MLE example for problem #2**

\( \mathcal{P} = \) distributions on \( n \) observations treated as iid \( N(\mu, \sigma^2) \) random variables.
- \( \Theta = \{ (\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0 \} \).
- Likelihood of \( (\mu, \sigma^2) \) given \( (Y_1, \ldots, Y_n) = (y_1, \ldots, y_n) \):
\[
\mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^{n} \phi_{\mu, \sigma^2}(y_i).
\]
- Often easier to determine maximizer of log-likelihood:
\[
\ln \mathcal{L}(\mu, \sigma^2) = \sum_{i=1}^{n} \ln \phi_{\mu, \sigma^2}(y_i) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 + \frac{n}{2} \ln \frac{1}{2\pi\sigma^2}.
\]
- Using calculus, we find that maximizing values of \( \mu \) and \( \sigma^2 \) are
\[
\hat{\mu}(y_1, \ldots, y_n) := \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \hat{\sigma}^2(y_1, \ldots, y_n) := \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu})^2.
\]
How good is this approach?

Again, consider $Y_1, \ldots, Y_n, Y$ iid random variables with $\mu := \mathbb{E}(Y)$.

1. We observe $Y_1, \ldots, Y_n$, and then form estimate

$$\hat{\mu}(Y_1, \ldots, Y_n) := \frac{1}{n} \sum_{i=1}^{n} Y_i.$$ 

2. We predict $\hat{y} = \hat{y}(Y_1, \ldots, Y_n) := \hat{\mu}(Y_1, \ldots, Y_n)$.

3. Outcome is $Y$, and squared loss is $(\hat{y} - Y)^2$.

A simple computation shows that, in expectation (over $Y_1, \ldots, Y_n$ and $Y$),

$$\mathbb{E}[(\hat{y}(Y_1, \ldots, Y_n) - Y)^2] = \left(1 + \frac{1}{n}\right) \text{var}(Y).$$

Recall: optimal prediction ($\hat{y} = \mu$) has $\mathbb{E}[(\hat{y}_o - Y)^2] = \text{var}(Y)$.

So, on average, using MLE is near-optimal when $n$ is large!

See assigned reading for details and discussion of problem #1.

Key takeaways

1. Statistical models for simple prediction problems, and the optimal predictions in these models.

2. How to derive near-optimal predictions from data in iid models.