CSE517A Machine Learning **Spring 2025**

Lecture 3: Estimating Probabilities from Data

Instructor: Marion Neumann

Reading: FCML 2.1-2.6 (Random Variables and Probability), 3.1-3.7 (Coin Game)

Learning Objective

Understand that many ML algorithms estimate probabilities. Appreciate that estimating probability distributions is beneficial. Get to know common estimators for the parameters of probability distributions.

Application

Imagine that we want to predict the **production of bushels of corn** per acre on a farm as a function of the proportion of that farm's planting area that was treated with a new pesticide. We have training data for 100 farms for this problem (https://www.developer.com/mgmt/real-world-machine-learning-model-evaluation-and-optimization.html).

http://www.corncapitalinnovations.com/production/300-bushel-corn/

Take some time to answer the following warm-up questions:
(1) *Is this a regression or classification problem?* (2) *What are the features?* (3) *What is the prediction task?* (4) *How well do you think a linear model will perform?*

1 Introduction

For general machine learning problem, our goal is to estimate the function f which satisfies $f(\mathbf{x}) = y$. For example, ordinary least squares regression wants to estimate the vector $\mathbf{w}$ which makes $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. However, there are some limitations: (1) y is only a *point estimate* (2) how do we deal with *noise*?

Therefore, it is a better idea to **estimate the conditional probability** $p(y \mid \mathbf{x})$. In this way, we can deal with uncertain outcomes/noise and incorporate prior knowledge.

1.1 Noise

Let's look at noise in our corn production application. Plotting the target (bushels of corn per acre) versus the feature (% of the farm treated) it is clear that an increasing, non-linear relationship exists, and that the data also have *random fluctuations*, cf. Figure 1[1].

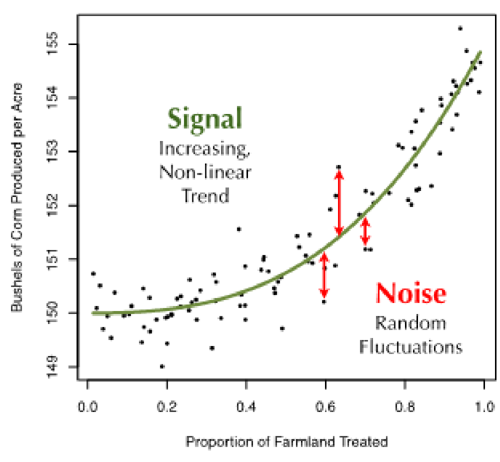

Figure 1: Signal-to-noise ratio in corn production application.

Exercise 1.1. How does noise look like in classification problems such as image classification?

[1] IMAGE SOURCE: https://www.developer.com/mgmt/real-world-machine-learning-model-evaluation-and-optimization.html

Discriminative vs. Generative Learning

Most supervised machine learning methods can be viewed as estimating $p(y \mid \mathbf{x})$ or $p(\mathbf{x}, y)$. When we estimate $p(y \mid \mathbf{x})$ directly, then we call it *discriminative learning*. When we estimate $p(\mathbf{x}, y) = p(\mathbf{x} \mid y)p(y)$, then we call it *generative learning*. In the following lectures, we will introduce examples for both.

1.2 Basic Problem: Tossing a Coin

Before we start thinking about estimating probability distributions in the context of regression and classification. Let's start with a simpler example. Imagine you find a funny looking coin and you start flipping it. Naturally, you ask yourself: *What is the probability that it comes up heads?*

You have the following data: H, T, T, H, H, H, T, T, T, T. What is $p(y = H)$ given that we observed n_H heads and n_T tails? So, intuitively,

$$p(y = H) = \frac{n_H}{n_H + n_T} = 0.4$$

Note: we have no $\mathbf{x}'s$ in this example. Let's formally derive this probability.

2 Maximum Likelihood Estimation

Let $p(y = H) = \theta$, where θ is the unknown parameter. All we have is D (sequence of heads and tails). So, the goal is to choose θ such that the observed data D is most likely.

Formally (MLE principle): Find $\hat{\theta}$ that maximizes the *likelihood of the data* $p(D \mid \theta)$:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} p(D \mid \theta) \tag{1}$$

For the sequence of coin flips we can focus on the number of successes in the sequence and use the binomial distribution (cf. FCML 2.3.2) to model $p(D \mid \theta)$[2]:

$$p(D \mid \theta) = \binom{n_H + n_T}{n_H} \theta^{n_H} (1 - \theta)^{n_T} = Bin(n_H \mid n, \theta) \tag{2}$$

Now,

$$\begin{aligned} \hat{\theta}_{\text{MLE}} &= \arg\max_{\theta} \binom{n_H + n_T}{n_H} \theta^{n_H} (1 - \theta)^{n_T} \\ &= \arg\max_{\theta} \log \binom{n_H + n_T}{n_H} + \log \theta^{n_H} + \log(1 - \theta)^{n_T} \\ &= \arg\max n_H \log \theta + n_T \log(1 - \theta) \end{aligned} \tag{3}$$

We can now solve for θ by taking the derivative and equating it to zero. This results in:

$$\frac{n_H}{\theta} = \frac{n_T}{1 - \theta} \Rightarrow n_H(1 - \theta) = n_T \theta \Rightarrow \hat{\theta}_{\text{MLE}} = \frac{n_H}{n_H + n_T} \tag{4}$$

Note that we found the (arg) maximum of the *log-likelihood* instead of the *likelihood*, which oftentimes leads to much easier to solve equations!

Advantages:

[2] Note that now technically our data is not the actual sequence, but only the number of times we see heads in the sequence, which is all we need for our scenario. We could also model the actual sequence; in that case we would omit the binomial coefficient in Eq. (2) altogether.

- MLE gives the explanation of the data you observed.
- If n is large and your model/distribution is correct (that is $\mathcal{H}$ includes the true model), then MLE finds the true parameters.

Disadvantages:

- But the MLE can overfit the data if n is small.
- If you do not have the correct model (and n is small) then MLE can be terribly wrong.

Exercise 2.1. What is the probability that my smartphone dies?
Let $y_1, \ldots, y_n \in \mathbb{R}$ with $y_i \geq 0$ be the customer-reported lifetimes of PEAR's popular smartphone JX. We further assume that lifetimes follow an exponential distribution:

$$p(y; \theta) = \theta e^{-\theta y}.$$

In order for you to use this distribution to compute the probability of your own JX phone dying next month, we need to estimate its parameter θ.

(a) Derive the log-likelihood $ll(\theta, y_1, \ldots, y_n)$.

(b) How do you derive the MLE estimator $\hat{\theta}$ based on $ll(\theta, y_1, \ldots, y_n)$? No computation required.

Exercise 2.2. Assume you model your data $y_1, \ldots, y_n$ with a Poisson distribution:

$$P(y; \theta) = \frac{\theta^y e^{-\theta}}{y!} \text{ for } y = 0, 1, 2, \ldots, K.$$

(a) Derive the negative log-likelihood (nll) of your data as a function of θ.

(b) For what data/events do you use a Poisson distribution? (You may search the internet for an answer.) Name a general definition and find at least two example applications.

3 The Bayesian Way and Maximum-a-posterior Estimation

For example, suppose you observe H,H,H,H,H. What is $\hat{\theta}_{\text{MLE}}$? Can we do something about this?

Answer: incorporate *prior knowledge*!

Say we think θ is close to q.
Simple fix (MAP smoothing): add m imaginary throws that would result in q.

For example, set $q = 0.5$, then add m heads and m tails to dataset D. Now,

$$\hat{\theta} = \frac{n_H + m}{n_H + n_T + 2m} \tag{5}$$

For large n, this change is insignificant; for small n, it incorporates your prior belief about θ. From Figure 2 we can see that MAP smoothing works well. But note that q is **uncertain**!

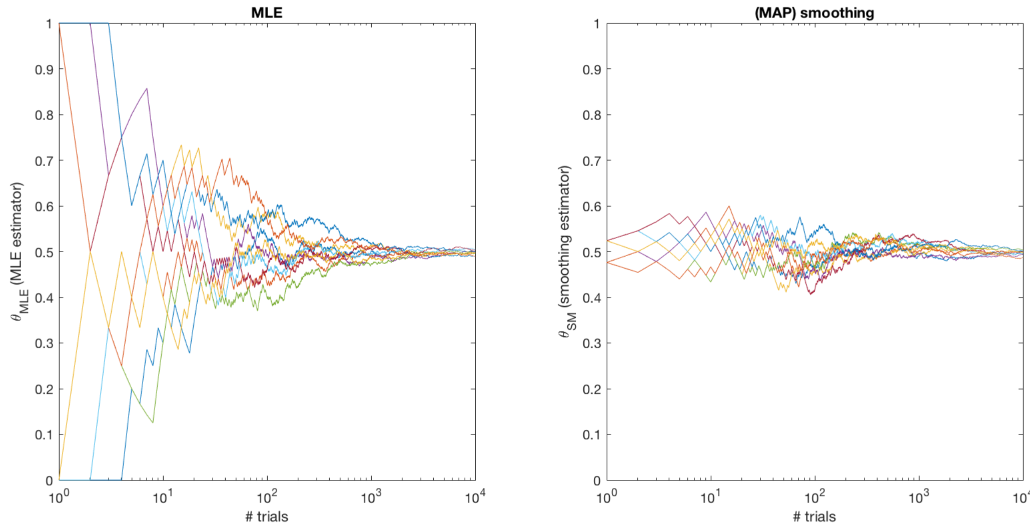

Figure 2: Comparing MLE to MAP smoothing we see that incorporating prior knowledge helps when observing few training examples

The *Bayesian way* is to model θ as a **random variable** drawn from a distribution $p(\theta)$. Note that θ is not a random variable associated with an event in a sample space. In frequentist statistics, this is forbidden. In Bayesian statistics, this is allowed.

Now, we can look at $p(\theta \mid D) = \frac{p(D|\theta)p(\theta)}{p(D)}$ (Bayes rule), where

- $p(D \mid \theta)$ is the **likelihood** of the data given the parameter(s) θ
- $p(\theta)$ is the **prior** distribution over the parameter(s) θ
- $p(\theta \mid D)$ is the **posterior** distribution over the parameter(s) θ

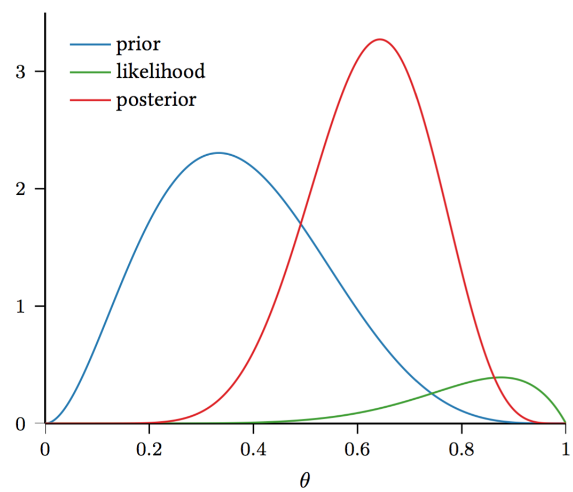

Figure 3: Prior distribution, likelihood, and posterior distribution.

θ is a **continuous univariate** RV on [0,1], we can use Beta distribution (cf. FCML 2.5.2) to model $p(\theta)$:

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{b(\alpha,\beta)} = Beta(\theta \mid \alpha, \beta) \tag{6}$$

where $b(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the *Beta function* that acts as a normalization constant. Note that here we only need a distribution over a univariate (1D) random variable. The multivariate generalization of the Beta distribution is the Dirichlet distribution.

Why do we use Beta distribution?

- it models continuous probabilities (θ lives on [0,1] and $\sum_i \theta = 1$)
- it is of the same distributional family as the binomial distribution (*conjugate prior*)
  $\rightarrow p(\theta \mid D) \propto p(D \mid \theta)p(\theta) \propto \theta^{n_H+\alpha-1}(1-\theta)^{n_T+\beta-1}$

Note:

$$p(\theta \mid D) = Beta(n_H + \alpha, n_T + \beta) \tag{7}$$

Note that in general θ are the parameters of our model. For the coin flipping scenario $\theta = p(y = H)$. So far, we have a distribution over θ. How can we get an estimate for θ?

For example, choose $\hat{\theta}$ to be the most likely θ given D.

Formally (MAP principle): Find $\hat{\theta}$ that maximizes the *posterior distribution over parameters* $p(\theta \mid D)$:

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \arg\max_{\theta} p(\theta \mid D) \\ &= \arg\max_{\theta} \frac{p(D \mid \theta)p(\theta)}{p(D)} \\ &= \arg\max_{\theta} \log p(D \mid \theta) + \log p(\theta)\end{aligned} \tag{8}$$

For coin flipping scenario, we get:

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \arg\max_{\theta} \log p(D \mid \theta) + \log p(\theta) \\ &= \arg\max_{\theta} n_H \log\theta + n_T \log(1-\theta) + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta) \\ &= \arg\max (n_H + \alpha - 1)\log\theta + (n_T + \beta - 1)\log(1-\theta)\end{aligned} \tag{9}$$

Solve for θ by taking the derivative and equating it to zero. This results in:

$$\hat{\theta}_{\text{MAP}} = \frac{n_H + \alpha - 1}{n_H + n_T + \alpha + \beta - 2} \tag{10}$$

Note that we found the (arg) maximum of the *log-posterior* instead of the *posterior*, which oftentimes leads to much easier to solve equations!

Advantages:

- as $n \rightarrow \infty$, $\hat{\theta}_{\text{MAP}} \rightarrow \hat{\theta}_{\text{MLE}}$
- MAP is a great estimator if prior belief exists and is accurate

Disadvantage:

- if n is small, it can be very wrong if prior belief is wrong
- also we have to choose a reasonable prior ($p(\theta) > 0 \ \forall\ \theta$)

4 "True" Bayesian Approach

MAP is ~~only~~ *one* way to get an estimator for θ. There is much more information in $p(\theta \mid D)$.

Posterior Mean

So, instead of the maximum as we did with MAP, we can use the posterior mean (end even its variance).

$$\hat{\theta}_{\text{MEAN}} = E[\theta, D] = \int_{\theta} \theta \, p(\theta \mid D) \, \mathrm{d}\theta \tag{11}$$

For coin flipping, this can be computed as $\hat{\theta}_{\text{MEAN}} = \frac{n_H + \alpha}{n_H + \alpha + n_T + \beta}$ as the posterior distribution is a Beta distribution, cf. Eq. 7, and the mean of $Beta(\alpha, \beta)$ is given by $\mu = \frac{\alpha}{\alpha + \beta}$.

For large n all three estimators (MLE, MAP, MEAN) will be the same, however, for a small number of observations we see that they are different as shown in Figure 4.

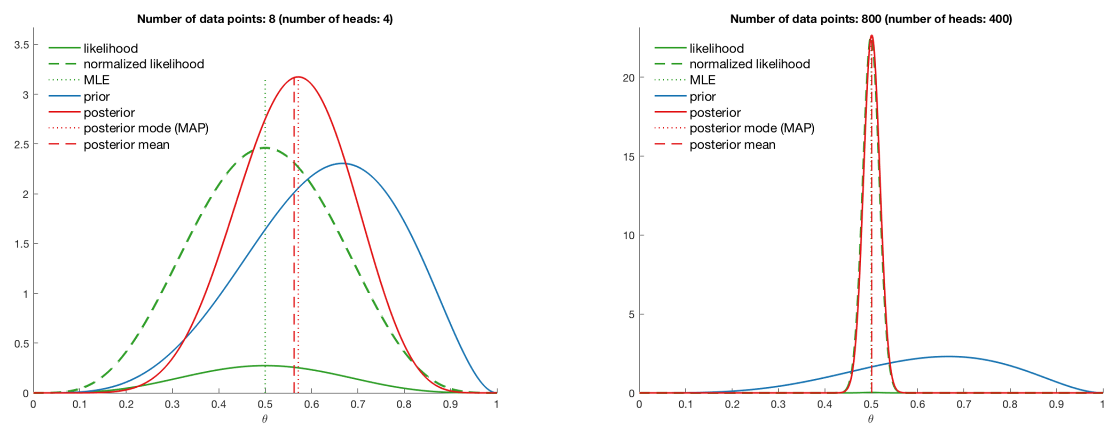

Figure 4: Probability density function (PDF) for likelihood, prior, and posterior (over parameters) and estimators (MLE, MAP, posterior mean) when the number of data points is small (*left*) and large (*right*).

Exercise 4.1. In our coin flipping example, show that $\hat{\theta}_{\text{MEAN}} \to \hat{\theta}_{\text{MAP}} \to \hat{\theta}_{\text{MLE}}$ as $n \to \infty$.

Posterior Predictive Distribution

So far, we talked about modeling and estimating parameters. But in Machine Learning we are actually interested in predictions. To directly estimate y from the given data, we can use the *posterior predictive distribution*. In our coin tossing example, this is given by:

$$\begin{aligned} p(y = H \mid D) &= \int_{\theta} p(y = H, \theta \mid D) \, \mathrm{d}\theta \\ &= \int_{\theta} p(y = H \mid \theta, D) \, p(\theta \mid D) \, \mathrm{d}\theta \\ &= \int_{\theta} \theta \, p(\theta \mid D) \, \mathrm{d}\theta \end{aligned} \tag{12}$$

Here, we used the fact that we defined $p(y = H) = \theta$ and that $p(y = H) = p(y = H \mid \theta, D)$ (this is only the case for coin flipping - not in general). Note that the prediction using the predictive posterior distribution is the same as $\hat{\theta}_{\text{MEAN}}$. Again, this nice result only holds for this particular example (coin flipping) and not in general.

5 Summary

List of concepts and terms to **understand** from this lecture:

- *data likelihood*
- *log-likelihood*
- *negative log-likelihood*
- *prior distribution over parameters*
- *posterior distribution over parameters*
- *posterior predictive distribution*
- MLE estimator
- MAP estimator

Exercise 5.1. Practice Retrieving!
For this summary exercise, it is intended that your answers are based on **your own** (current) understanding of the concepts (and not on the definitions you read and copy from these notes or from elsewhere). Don't hesitate to **say it out loud** to your seat neighbor, your pet or stuffed animal, or to yourself before **writing it down**. Research studies show that this practice of retrieval and phrasing out loud will help you retain the knowledge!

(a) Using *your own words*, summarize each of the concepts listed above in 2-3 sentences by retrieving the knowledge from the top of your head.

(b) Why are we interested in the log likelihood?

(c) What is the main difference of the MLE estimator and MAP estimator?

(d) What is the difference between the *posterior distribution over parameters* and the *posterior predictive distribution*?

(e) Why do we need the *posterior distribution over parameters*? Give a guess now. We will cover this in the next lecture!

And always remember: It's not bad to get it wrong. *Getting it wrong is part of learning!* Use your notes or other resources to get the correct answer or come to our office hours to get help!