CS 412 — Introduction to Machine Learning (UIC)	February 08, 2025

Lecture 6

Lecturer: Aadirupa Saha *Scribe(s): Yugesh Sappidi*

1 Overview

In the last lecture, we covered the following main topics:

- Continuation of Multiclass Logistic Regression
- Maximum Likelihood Estimation (MLE) for Bernoulli.
- Maximum *a posteriori* (MAP) Estimation for Bernoulli.

This lecture focuses on:

- Maximum Likelihood Estimation (MLE) for Categorical, Logistic, and Linear regression.
- Maximum *a posteriori* (MAP) estimation and its interpretation as regularization for Linear Regression.
- Regularized linear regression using L2 (Gaussian prior) and L1 (Laplace prior) penalties.

2 Recap of Last lecture

2.1 Bernoulli Distribution: MLE and MAP

Consider a binary random variable

$$y \sim \mathrm{Ber}(\mu)$$

with unknown parameter $\mu \in [1]$. Suppose we have a dataset

$$D = \{y_1, \ldots, y_n\}$$

where each $y_i \sim \mathrm{Ber}(\mu)$ independently.

MLE for the Bernoulli Distribution. The maximum likelihood estimate (MLE) is found by maximizing the likelihood function:

$$L(\mu; D) = \prod_{i=1}^{n} \mu^{y_i}(1-\mu)^{1-y_i}.$$

Let

$$N_1 = \#\{i : y_i = 1\} \quad \text{and} \quad N_0 = n - N_1.$$

Then the MLE is given by

$$\hat{\mu}_{\text{MLE}} = \frac{N_1}{n}.$$

MAP with a Beta Prior. Assume a conjugate Beta prior for μ:

$$\mu \sim \mathrm{Beta}(a, b).$$

Then the posterior distribution is also Beta, and the MAP estimate becomes

$$\hat{\mu}_{\text{MAP}} = \frac{N_1 + (a-1)}{n + (a+b-2)}.$$

For example, if a $\mathrm{Beta}(1,1)$ (uniform) prior is assumed, then:

$$\hat{\mu}_{\text{MAP}} = \frac{N_1}{n} \quad \text{(i.e. the MAP equals the MLE).}$$

The extra pseudo-counts $(a-1)$ and $(b-1)$ act as a regularizer for extreme cases.

3 Multiclass Classification and Categorical Distributions

Suppose we now have a multiclass classification task. Let the labels be

$$y \in \{1, 2, \ldots, C\}.$$

We assume that each label follows a categorical distribution:

$$y \sim \text{Cat}(p_1, p_2, \ldots, p_C)$$

with probabilities satisfying

$$p_j \in [1], \quad \sum_{j=1}^{C} p_j = 1.$$

3.1 MLE for the Categorical Distribution

Given a dataset

$$D = \{y_1, \ldots, y_n\},$$

let

$$N_j = \#\{i : y_i = j\} \quad \text{for } j = 1, \ldots, C.$$

The likelihood function is

$$L(p; D) = \prod_{i=1}^{n} \prod_{j=1}^{C} p_j^{\mathbb{I}(y_i = j)},$$

and by taking the log-likelihood we have

$$\log L(p; D) = \sum_{j=1}^{C} N_j \log p_j.$$

Maximizing this subject to $\sum_{j=1}^{C} p_j = 1$ yields the MLE:

$$\hat{p}_j^{\text{MLE}} = \frac{N_j}{n}.$$

A corresponding Bayesian/MAP approach would introduce a Dirichlet prior (the multivariate generalization of the Beta distribution) on $p = (p_1, \ldots, p_C)$.

4 Logistic Regression as an MLE Problem

Logistic regression can be viewed as a generalization of the categorical MLE. Suppose that each feature vector $x \in \mathbb{R}^d$ (for example, representing a flower by its measurements) is associated with a label $y \in \{1, 2, \ldots, C\}$. In logistic regression, we parameterize the class probabilities as:

$$P(y = j \mid x; W) = \frac{e^{\omega_j^\top x}}{\sum_{k=1}^{C} e^{\omega_k^\top x}},$$

where

$$W = [\omega_1, \ldots, \omega_C] \in \mathbb{R}^{d \times C}.$$

4.1 Negative Log-Likelihood (Cross-Entropy Loss)

Given a dataset $D = \{(x_i, y_i)\}_{i=1}^n$, the likelihood is:

$$L(W; D) = \prod_{i=1}^{n} P(y_i \mid x_i; W).$$

Taking the negative log, we obtain the loss function:

$$-\log L(W; D) = \sum_{i=1}^{n} [-\log P(y_i \mid x_i; W)] = \sum_{i=1}^{n} \left[\log \sum_{k=1}^{C} e^{\omega_k^\top x_i} - \omega_{y_i}^\top x_i \right].$$

This is minimized with respect to the parameter matrix W during training.

Numerical Stability. In practice, the log-sum-exp trick is used:

$$\log \sum_{k=1}^{C} e^{\omega_k^\top x_i} = \omega_{\text{max}}^\top x_i + \log \sum_{k=1}^{C} e^{(\omega_k - \omega_{\text{max}})^\top x_i},$$

where ω_{max} is chosen so that the exponentials remain numerically stable.

5 Linear Regression Through MLE

Consider a dataset with inputs and outputs:

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

Assume that the outputs are generated by a linear model plus Gaussian noise:

$$y_i \mid x_i \sim \mathcal{N}(w^\top x_i, \sigma^2).$$

5.1 Deriving the MLE

The likelihood function is:

$$L(w; D) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - w^\top x_i)^2}{2\sigma^2} \right).$$

Taking the logarithm, we have (ignoring constants):

$$\log L(w; D) \propto -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^\top x_i)^2.$$

Maximizing this likelihood is equivalent to minimizing the sum of squared errors:

$$\hat{w}_{\text{MLE}} = \arg\min_w \sum_{i=1}^{n} (y_i - w^\top x_i)^2.$$

Variance Estimation. The MLE for σ^2 is:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - w^\top x_i)^2,$$

although for an unbiased estimate one would use $n - d$ in the denominator, where d is the number of features.

6 Regularization as MAP Estimation

Regularization in linear regression can be interpreted in a Bayesian setting via MAP estimation, where the regularization term corresponds to the log-prior.

6.1 L2 Regularization (Ridge Regression)

Assume a Gaussian prior on w:

$$w \sim \mathcal{N}(0, \lambda^{-1} I).$$

The log-prior is proportional to $-\lambda \|w\|_2^2$. Combining this with the likelihood gives the MAP estimate:

$$\hat{w}_{\text{MAP}} = \arg\min_w \left[\sum_{i=1}^{n} (y_i - w^\top x_i)^2 + \lambda \|w\|_2^2 \right].$$

In closed form (when applicable), the solution is:

$$\hat{w} = \left(X^\top X + \lambda I \right)^{-1} X^\top y,$$

which can also be interpreted as adding λ to the eigenvalues of $X^\top X$ for numerical stability.

6.2 L1 Regularization (Lasso Regression)

Assume a Laplace prior on each component w_j:

$$p(w_j) \propto \exp\left(-\frac{|w_j|}{\lambda}\right).$$

Thus, the MAP estimate becomes:

$$\hat{w}_{\text{MAP}} = \arg\min_w \left[\sum_{i=1}^{n}(y_i - w^\top x_i)^2 + \lambda\|w\|_1\right].$$

The L1 penalty tends to produce sparse solutions by forcing some coefficients exactly to zero.

7 Advanced Topics and Practical Considerations

7.1 Hierarchical Models and Hyperparameter Tuning

The lecture also touches upon hierarchical Bayesian models where hyperparameters (such as a, b in the Beta prior or λ in the Gaussian prior) can be inferred from the data rather than set manually. One common method is the **Empirical Bayes** approach, where hyperparameters are estimated by maximizing the marginal likelihood:

$$\lambda^* = \arg\max_\lambda \int P(D \mid w)P(w \mid \lambda)\, dw.$$

7.2 Limitations and Considerations

- **Overfitting:** MLE without regularization may lead to overfitting, especially in high-dimensional settings.
- **Prior Sensitivity:** MAP estimates can be sensitive to the choice of prior. A misspecified prior (e.g., Laplace when the true coefficients are not sparse) can hurt performance.
- **Computational Complexity:** Bayesian approaches with non-conjugate priors may require approximate inference techniques (such as Markov Chain Monte Carlo or variational inference).

8 Summary

In this lecture we unified frequentist and Bayesian approaches in machine learning:

- MLE provides a framework where loss minimization (e.g., cross-entropy for logistic regression or squared loss for linear regression) emerges naturally from the assumption of a data likelihood.
- MAP estimation shows that regularization is equivalent to incorporating prior beliefs about the parameters.
- Logistic regression and linear regression can both be derived from probabilistic assumptions on the data, linking optimization to probability theory.

Future topics may extend these ideas to more advanced models such as non-conjugate priors and hierarchical models.

References:

1. Probabilistic Machine Learning: An Introduction (PML), Sections 4.2.5 - 4.2.7, 4.5
2. Lecture notes by Prof. Saha Adhirupa from course CS 412 Lec 6