

Lec 9: Logistic Regression (CS3780/5780, Sp26)

1 Naive Bayes Recap

The key assumption for the Naive Bayes model is that conditioned on the label $Y = y$, the various features of input are independent of each other. That is,

$$P(X = \mathbf{x} | Y = y) = \prod_{\alpha=1}^d P(X[\alpha] = \mathbf{x}[\alpha] | Y = y)$$

So if we think about the spam classification example, our assumption says, conditioned on an email being spam (or not spam) the draw of various words from the english dictionary is independent of each other. This assumption is often not true but the model simplifies drastically and gives rise to simple but surprisingly effective models in practice.

The power of Naive Bayes algorithm comes from its versatility. In the same model we could have some features to be binary, some to be categorical and yet others continuous. Since we are modeling each feature conditioned on its label separately we can model them all in differently under the same hood. Further when it comes time to predict the label y of a given instance \mathbf{x} we simply use Bayes rule to get

$$P(Y = y | X = \mathbf{x}) = \frac{\prod_{\alpha=1}^d P(X[\alpha] = \mathbf{x}[\alpha] | Y = y) P(Y = y)}{\sum_{y' \in \mathcal{Y}} \prod_{\alpha=1}^d P(X[\alpha] = \mathbf{x}[\alpha] | Y = y') P(Y = y')}$$

1.1 Examples: Categorical Variables

Categorical features are one where $\mathbf{x}[\alpha]$ can take on exactly one of a finite set of values. Example say feature $\alpha = 1$ is such that $\mathbf{x}[1] \in \{CA, NY, PA, MA\}$. In this case, for each feature we can parameterize it by a probability vector θ_α whose dimension is the same as number of values that coordinate can take. In this case,

$$P(X[\alpha] = k | Y = y) = \theta_{\alpha,+1}[k]$$

1.2 Examples: Multinomial Variables

If on the other hand, features don't represent categories but counts, we need to use a different model. E.g. in the document categorization, feature value $x[\alpha] = j$ means that in this particular document, the α 'th word in my dictionary appears j times. In this case one can use the multinomial distribution where θ is a distribution over all the d words in the dictionary and

$$P(X = \mathbf{x} | Y = y) = \frac{m!}{\prod_{\alpha=1}^d x[\alpha]!} \prod_{\alpha=1}^d (\theta_y[\alpha])^{x[\alpha]} \propto \prod_{\alpha=1}^d (\theta_y[\alpha])^{x[\alpha]}$$

where m is the number of words in the document.

1.3 Examples: Continuous Variables

Finally for continuous variables one could model each feature conditioned on class label as a uni-variate gaussian as

$$P(X[\alpha] = \mathbf{x}[\alpha]|Y = y) = \frac{1}{\sqrt{2\pi\sigma_{\alpha,y}^2}} \exp\left(-\frac{1}{2\sigma_{\alpha,y}^2}(x[\alpha] - \mu_\alpha)^2\right)$$

An interesting special case of gaussian NB model is one where for each coordinate we fix variance to be same across classes, or in other words have $\sigma_{\alpha,y}^2 = \sigma_\alpha^2$ for every $y \in \mathcal{Y}$

2 Naive Bayes and the Logistic Model

2.1 Multinomial NB

$$\begin{aligned} P(Y = 1|X = \mathbf{x}) &= \frac{\prod_{\alpha=1}^d P(X[\alpha] = \mathbf{x}[\alpha]|Y = 1)P(Y = 1)}{\prod_{\alpha=1}^d P(X[\alpha] = \mathbf{x}[\alpha]|Y = 1)P(Y = 1) + \prod_{\alpha=1}^d P(X[\alpha] = \mathbf{x}[\alpha]|Y = -1)P(Y = -1)} \\ &= \frac{\prod_{\alpha=1}^d \theta_{\alpha,+1}^{\mathbf{x}[\alpha]} P(Y = 1)}{\prod_{\alpha=1}^d \theta_{\alpha,+1}^{\mathbf{x}[\alpha]} P(Y = 1) + \prod_{\alpha=1}^d \theta_{\alpha,-1}^{\mathbf{x}[\alpha]} P(Y = -1)} \end{aligned}$$

Set $\mathbf{w}_{+1}[\alpha] = \log(\theta_{\alpha,+1})$, $\mathbf{w}_{-1}[\alpha] = \log(\theta_{\alpha,-1})$ and set $b_{+1} = \log(P(Y = 1))$ and $b_{-1} = \log(P(Y = -1))$ so that

$$\begin{aligned} P(Y = 1|X = \mathbf{x}) &= \frac{\prod_{\alpha=1}^d \exp(\mathbf{w}_{+1}[\alpha] \cdot \mathbf{x}[\alpha]) \times \exp(b_{+1})}{\prod_{\alpha=1}^d \exp(\mathbf{w}_{+1}[\alpha] \cdot \mathbf{x}[\alpha]) \times \exp(b_{+1}) + \prod_{\alpha=1}^d \exp(\mathbf{w}_{-1}[\alpha] \cdot \mathbf{x}[\alpha]) \times \exp(b_{-1})} \\ &= \frac{\exp\left(\sum_{\alpha=1}^d \mathbf{w}_{+1}[\alpha] \cdot \mathbf{x}[\alpha] + b_{+1}\right)}{\exp\left(\sum_{\alpha=1}^d \mathbf{w}_{+1}[\alpha] \cdot \mathbf{x}[\alpha] + b_{+1}\right) + \exp\left(\sum_{\alpha=1}^d \mathbf{w}_{-1}[\alpha] \cdot \mathbf{x}[\alpha] + b_{-1}\right)} \\ &= \frac{1}{1 + \exp\left(\sum_{\alpha=1}^d (\mathbf{w}_{-1}[\alpha] - \mathbf{w}_{+1}[\alpha]) \cdot \mathbf{x}[\alpha] + b_{-1} - b_{+1}\right)} \\ &= \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x} - b)} \end{aligned}$$

where $\mathbf{w} = \mathbf{w}_{+1} - \mathbf{w}_{-1}$ and $b = b_{+1} - b_{-1}$.

On similar lines as the above, one can also show that for the gaussian Naive Bayes model, with variance across the class in each feature α is fixed to σ_α^2 , one again gets the conclusion that $P(Y = 1|X = \mathbf{x})$ can be written in the form $\frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x} - b)}$ for appropriate \mathbf{w}, b . Such a model of the conditional distribution $P(Y = 1|X = \mathbf{x})$ is called the Logistic model.

3 Discriminative Model: Logistic Regression

In Naive Bayes, we model $P(X, Y)$ jointly but in the end while trying to predict label given \mathbf{x} , we only need to evaluate $P(Y|X)$ using the Naive Bayes assumption. For the multinomial and Categorical cases and for fixed variance across class Gaussian Naive Bayes models, we have that $P(Y = 1|X)$ follows a logistic model. The idea behind Logistic regression is to only model $P(Y|X)$

directly using Logistic model and not bothering to model the joint $P(X, Y)$. Such an approach is called Discriminative model. Unlike Naive Bayes like approaches that are called generative models which model the joint, for discriminative model, we only model $P(Y|X)$ (in logistic regression case using Logistic model) and don't make any assumptions of $P(X)$. Specifically for Logistic regression we make the assumption that

$$P(Y = 1|X = \mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{w}^\top \mathbf{x} + b))} = \sigma(\mathbf{w}^\top \mathbf{x} + b)$$

where $\sigma(x) = \frac{1}{1 + \exp(-x)}$ is the sigmoid function. Notice that $\sigma(0) = 1/2$ and so for logistic regression, the classification boundary is the halfspace given by $\text{sign}(\mathbf{w}^\top \mathbf{x} + b)$.

4 MLE and MAP for Logistic Regression

$$\begin{aligned} \hat{\mathbf{w}}_{\text{MLE}} &= \arg \max_{\mathbf{w}} P(D|\mathbf{w}) && \text{(Defn. of MLE)} \\ &= \arg \max_{\mathbf{w}} P((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)|\mathbf{w}) && \text{(Substituting for D)} \\ &= \arg \max_{\mathbf{w}} \prod_{i=1}^n P((\mathbf{x}_i, y_i)|\mathbf{w}) && \text{(i.i.d. assumption)} \\ &= \arg \max_{\mathbf{w}} \prod_{i=1}^n P(y_i|\mathbf{x}_i, \mathbf{w}) \times P(\mathbf{x}_i|\mathbf{w}) && \text{(chain rule)} \\ &= \arg \max_{\mathbf{w}} \prod_{i=1}^n P(y_i|\mathbf{x}_i, \mathbf{w}) \times P(\mathbf{x}_i) && \text{(P(X) doesn't depend on w)} \\ &= \arg \max_{\mathbf{w}} \prod_{i=1}^n P(y_i|\mathbf{x}_i, \mathbf{w}) && \text{(dropping product with positive constant)} \\ &= \arg \max_{\mathbf{w}} \sum_{i=1}^n \log(P(y_i|\mathbf{x}_i, \mathbf{w})) && \text{(taking log)} \\ &= \arg \max_{\mathbf{w}} \sum_{i=1}^n \log\left(\frac{1}{1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)}\right) && \text{(substituting logistic form)} \\ &= \arg \min_{\mathbf{w}} \sum_{i=1}^n \log\left(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)\right) && \text{(maximizing function is same as minimizing the negative)} \end{aligned}$$

If instead of MLE we assume a prior on model parameter \mathbf{w} as $P(\mathbf{w})$ then we can do a MAP

estimate on similar lines as:

$$\begin{aligned}
\hat{\mathbf{w}}_{\text{MAP}} &= \arg \max_{\mathbf{w}} P(\mathbf{w}|D) \\
&= \arg \max_{\mathbf{w}} P(D|\mathbf{w}) \times P(\mathbf{w}) \\
&= \arg \max_{\mathbf{w}} P((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)|\mathbf{w}) \times P(\mathbf{w}) \\
&= \arg \max_{\mathbf{w}} \prod_{i=1}^n P((\mathbf{x}_i, y_i)|\mathbf{w}) \times P(\mathbf{w}) \\
&= \arg \max_{\mathbf{w}} \sum_{i=1}^n \log(P(y_i|\mathbf{x}_i, \mathbf{w})) + \log(P(\mathbf{w})) \\
&= \arg \min_{\mathbf{w}} \sum_{i=1}^n \log\left(1 + \exp\left(-y_i \mathbf{w}^\top \mathbf{x}_i\right)\right) - \log(P(\mathbf{w}))
\end{aligned}$$

If we consider as an example the prior

$$P(\mathbf{w} \propto \exp\left(-\frac{\mathbf{w}^\top \mathbf{w}}{2\sigma^2}\right))$$

then we have:

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg \min_{\mathbf{w}} \sum_{i=1}^n \log\left(1 + \exp\left(-y_i \mathbf{w}^\top \mathbf{x}_i\right)\right) + \frac{1}{2\sigma^2} \mathbf{w}^\top \mathbf{w}$$

5 Summary

Given an instance \mathbf{x}, y and model weights \mathbf{w} , one can view the negative log of $P(Y = y|X = \mathbf{x})$ given by

$$\ell(\mathbf{w}^\top \mathbf{x}, y) = \log\left(1 + \exp\left(-y \mathbf{w}^\top \mathbf{x}\right)\right)$$

as a loss function often termed the logistic loss. One can view the hypothesis as linear hypothesis $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. Now the one question left for us to answer is, if we are interested in either the MLE (or the MAP) solution given by

$$\hat{\mathbf{w}}_{\text{MLE}} = \arg \min_{\mathbf{w}} \sum_{i=1}^n \log\left(1 + \exp\left(-y_i \mathbf{w}^\top \mathbf{x}_i\right)\right)$$

one needs to figure out a way to optimize the above function w.r.t. \mathbf{w} . This is easier said than done. Trying to find an exact solution by setting gradient to 0 and solving for \mathbf{w} is not a viable option as the equation to solve is very complex. In the next couple lectures we will touch upon optimization techniques like gradient descent, newtons method and stochastic gradient descent techniques to solve the optimization using iterative methods.

Finally its worth noting that one can extend the logistic model to multiclass setting using

$$P(Y = k|X = \mathbf{x}) = \frac{\exp(-\mathbf{w}_k^\top \mathbf{x})}{\sum_{j=1}^K \exp(-\mathbf{w}_j^\top \mathbf{x})}$$

for K parameters $\mathbf{w}_1, \dots, \mathbf{w}_K$ each in d dimensions.