

1 Definitions Ahoy!

1.1 MLE/MAP

1. **MLE:** Finds the best parameters for a specific dataset, $\mathcal{D}$. Specifically, we want to find the parameters $\hat{\theta}_{MLE}$ that maximize the likelihood for $\mathcal{D}$.

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta} p(\mathcal{D} \mid \theta)$$

2. **MAP:** Finds the best parameters given $\mathcal{D}$ and a prior belief about the parameters. Specifically, we want to find the parameters $\hat{\theta}_{MAP}$ that maximize the posterior distribution $p(\theta \mid \mathcal{D})$ of parameters θ .

$$\begin{aligned} \hat{\theta}_{MAP} &= \operatorname{argmax}_{\theta} p(\theta \mid \mathcal{D}) \\ &= \operatorname{argmax}_{\theta} \frac{p(\mathcal{D} \mid \theta)p(\theta)}{\text{Normalizing Constant}} \\ &= \operatorname{argmax}_{\theta} p(\mathcal{D} \mid \theta)p(\theta) \\ &= \operatorname{argmin}_{\theta} -\log(p(\mathcal{D} \mid \theta)p(\theta)) \\ &= \operatorname{argmin}_{\theta} -\log p(\mathcal{D} \mid \theta) - \log p(\theta) \end{aligned}$$

3. **MLE and MAP for conditional likelihood:** When we want to predict the output y given the input x using our supervised dataset, we have to reformulate the MLE and MAP optimizations to use the conditional likelihood (and conditional posterior) instead:

$$\begin{aligned} \hat{\theta}_{MLE} &= \operatorname{argmax}_{\theta} p(\mathcal{D} \mid \theta) \\ &= \operatorname{argmax}_{\theta} \prod_{i=1}^N p(y^{(i)} \mid x^{(i)}, \theta) \\ &= \operatorname{argmin}_{\theta} -\log \prod_{i=1}^N p(y^{(i)} \mid x^{(i)}, \theta) \\ &= \operatorname{argmin}_{\theta} -\sum_{i=1}^N \log p(y^{(i)} \mid x^{(i)}, \theta) \end{aligned}$$

$$\begin{aligned} \hat{\theta}_{MAP} &= \operatorname{argmax}_{\theta} p(\theta \mid \mathcal{D}) \\ &= \operatorname{argmax}_{\theta} \left(\prod_{i=1}^N p(y^{(i)} \mid x^{(i)}, \theta) \right) p(\theta) \\ &= \operatorname{argmin}_{\theta} -\log \prod_{i=1}^N p(y^{(i)} \mid x^{(i)}, \theta) - \log p(\theta) \\ &= \operatorname{argmin}_{\theta} -\sum_{i=1}^N \log p(y^{(i)} \mid x^{(i)}, \theta) - \log p(\theta) \end{aligned}$$

2 Anybody have a MAP?

Imagine you are a data scientist working for an advertising company. The advertising company has recently run an ad and they want you to estimate its performance. The ad was shown to N people. $y^{(i)} = 1$ if person i clicked on the ad and 0 otherwise. Thus $\sum_i^N y^{(i)} = N_1$ people decided to click on the ad. Assume that the probability that the i -th person clicks on the ad is ϕ and the probability that the i -th person does not click on the ad is $1 - \phi$.

Note

$$p(\mathcal{D} \mid \theta) = p\left(y^{(1)}, y^{(2)}, \dots, y^{(N)} \mid \phi\right) = \prod_{i=1}^N p\left(y^{(i)} \mid \phi\right) = \phi^{N_1} (1 - \phi)^{N - N_1}$$

1. Calculate $\hat{\phi}_{MLE}$.

$$\begin{aligned}\hat{\phi}_{MLE} &= \underset{\phi}{\operatorname{argmin}} -\log(p(\mathcal{D} \mid \phi)) \\ &= \underset{\phi}{\operatorname{argmin}} -\log(\phi^{N_1} (1 - \phi)^{N - N_1}) \\ &= \underset{\phi}{\operatorname{argmin}} -N_1 \log(\phi) - (N - N_1) \log(1 - \phi)\end{aligned}$$

Setting the derivative equal to zero yields

$$\begin{aligned}0 &= \frac{-N_1}{\phi} + \frac{(N - N_1)}{1 - \phi} \\ \implies \phi_{MLE} &= \frac{N_1}{N}\end{aligned}$$

2. Your coworker tells you that $\phi \sim \text{Beta}(\alpha, \beta)$. That is:

$$p(\phi) = \frac{\phi^{\alpha-1} (1 - \phi)^{\beta-1}}{B(\alpha, \beta)}$$

Note that $B(\alpha, \beta)$ is not a function of ϕ and can be treated as a constant. Formulate the optimization of the log posterior, $\underset{\phi}{\operatorname{argmin}} -\log p(\phi \mid \mathcal{D})$, in terms of N, N_1, ϕ, α , and β .

$$\begin{aligned}&\underset{\phi}{\operatorname{argmin}} -\log p(\phi \mid \mathcal{D}) \\ &= \underset{\phi}{\operatorname{argmin}} -\log(p(\mathcal{D} \mid \phi)p(\phi)) \quad \text{By Bayes rule and dropping } 1/p(\mathcal{D}) \\ &= \underset{\phi}{\operatorname{argmin}} -\log p(\mathcal{D} \mid \phi) - \log p(\phi) \\ &= \underset{\phi}{\operatorname{argmin}} -\log(\phi^{N_1} (1 - \phi)^{N - N_1}) - \log \frac{\phi^{\alpha-1} (1 - \phi)^{\beta-1}}{B(\alpha, \beta)} \\ &= \underset{\phi}{\operatorname{argmin}} -\log \phi^{N_1} - \log(1 - \phi)^{N - N_1} - \log \phi^{\alpha-1} - \log(1 - \phi)^{\beta-1} \quad \text{Drop } B(\alpha, \beta) \\ &= \underset{\phi}{\operatorname{argmin}} -N_1 \log \phi - (N - N_1) \log(1 - \phi) - (\alpha - 1) \log \phi - (\beta - 1) \log(1 - \phi) \\ &= \underset{\phi}{\operatorname{argmin}} -(N_1 + \alpha - 1) \log \phi - (N - N_1 + \beta - 1) \log(1 - \phi)\end{aligned}$$

3. Now, calculate $\hat{\phi}_{MAP}$.

$$\begin{aligned}\hat{\phi}_{MAP} &= \operatorname{argmin}_{\phi} -\log(p(\mathcal{D} \mid \phi)p(\phi)) \\ &= \operatorname{argmin}_{\phi} -(N_1 + \alpha - 1)\log(\phi) - (N - N_1 + \beta - 1)\log(1 - \phi)\end{aligned}$$

Setting the derivative equal to zero yields

$$\begin{aligned}0 &= \frac{-N_1 - \alpha + 1}{\phi} + \frac{(N - N_1 + \beta - 1)}{1 - \phi} \\ \implies \phi_{MAP} &= \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2}\end{aligned}$$

4. Suppose $N = 100$ and $N_1 = 10$. Furthermore, you believe that in general people click on ads about 6 percent of the time, so you, somewhat naively, decide to set $\alpha = 6 + 1 = 7$, and $\beta = 100 - 6 + 1 = 95$. Calculate $\hat{\phi}_{MAP}$.

$$\hat{\phi}_{MAP} = \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2} = \frac{10 + 7 - 1}{100 + 102 - 2} = \frac{16}{200} = 0.08$$

5. How do $\hat{\phi}_{MLE}$ and $\hat{\phi}_{MAP}$ differ? Argue which estimate you think is better.

Both estimates are reasonable given the available information. If you believe that this advertisement is similar to those advertisements that averaged a 6 percent click rate, then $\hat{\phi}_{MAP}$ may be a superior estimate, but if the circumstances under which the advertisement was shown were different from the usual, then $\hat{\phi}_{MLE}$ might be a better choice.

3 Conceptual MLE/MAP Questions

1. When calculating the MAP estimates, we rely on the Bayes formula and then argue we can ignore $p(\mathcal{D})$. Why do we usually ignore calculating $p(\mathcal{D})$?

We are taking the argmax over θ and since the value of $P(\mathcal{D})$ is not related to θ we can ignore this value.

2. As the amount of data increases, how are MLE and MAP affected?

MLE finds the θ that maximizes the likelihood term $p(\mathcal{D}|\theta)$. $p(\mathcal{D}|\theta)$ becomes more informative about the true parameter values because there is more data.

MAP finds the θ that maximizes the posterior, $p(\theta | \mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$. The influence of the prior term $p(\theta)$ decreases relative to the likelihood term because the prior becomes less significant in comparison to the increasing amount of data.

3. Can MLE and MAP estimates be the same? If so, when?

As the amount of data increases, the MLE and MAP estimates converge to the same value. Therefore, with enough data, MLE and MAP will be approximately equal.

Additionally, if MAP uses a uniform distribution as the prior, the MLE and MAP estimates will be equal. In other words, you can think of MLE as a special case of MAP, where the prior is uniform!

4 Let's Generative Some Definitions!

Generative Models: class of models that can generate new data points that are similar to the training data. That is, generative models model $P(Y | \theta_{\text{class}})$ and $P(X | Y, \theta_{\text{class conditional}})$. In addition to generating new data, generative models can also be used for prediction.

The parameters θ_{class} and $\theta_{\text{class conditional}}$ are learned from the data. Bayes rule can then be used to compute $P(Y | X, \theta)$

$$P(Y | X, \theta_{\text{class conditional}}, \theta_{\text{class}}) \propto P(X | Y, \theta_{\text{class conditional}}) P(Y | \theta_{\text{class}})$$

Properties:

- More model assumptions involved
- Rich model allows for generating new data points that are similar to the training data.
- Can be used for unsupervised learning, where the class labels are not known.

Discriminative Models: Discriminative models model $P(Y | X, \theta)$ directly. The parameters, θ , are also learned from the data.

Properties:

- Less model assumptions needed
- Requires more labeled data

4.1 Generative + MAP

We can also combine generative models with MAP by adding a prior on the parameters:

$$P(Y | X, \theta) P(\theta) \propto P(X | Y, \theta_{\text{class conditional}}) P(Y | \theta_{\text{class}}) P(\theta_{\text{class conditional}}) P(\theta_{\text{class}})$$

5 Comparing models

	MLE	MAP
Discriminative	• Linear Regression • Logistic Regression • Logistic Regression with polynomial features	• Linear regression with L2 regularization • Logistic Regression with Laplace Prior
Generative	• Naive Bayes	• Naive Bayes with Laplace smoothing*

Consider the list of models we've learned so far and place them in the correct boxes above:

- Linear regression
- Logistic regression
- Linear regression with L2 regularization
- Logistic regression with Laplace prior
- Logistic regression with polynomial features
- Naive Bayes

5.1 Reminder: Regularization and MAP

	Regularization Penalty	Prior
Ridge Regression	$\ \mathbf{w}\ _2^2$	$w_j \sim \mathcal{N}(0, \tau^2)$
Lasso	$\ \mathbf{w}\ _1$	$w_j \sim \text{Laplace}(0, b)$

6 Gaussian Discriminant Analysis (A generative method)

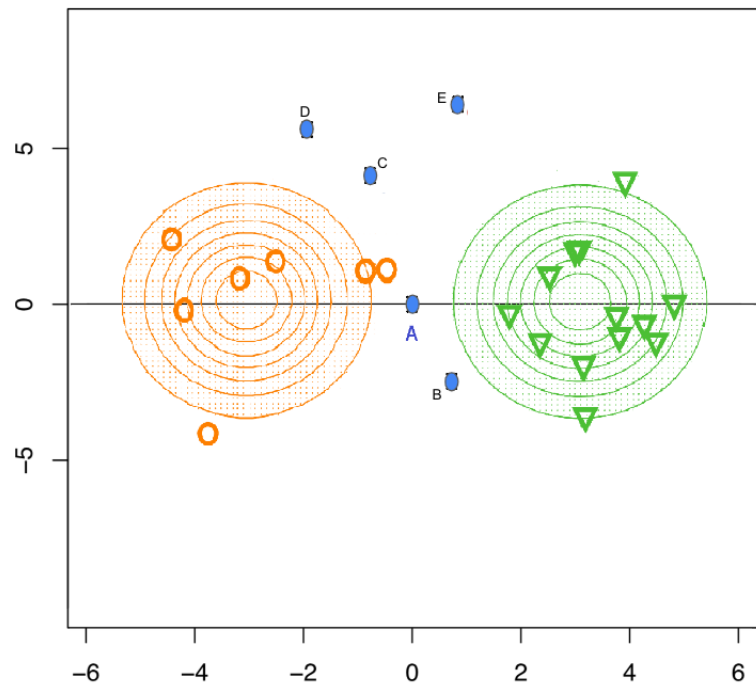
Gaussian discriminant analysis is used when the input features are continuous and $p(\mathbf{x} | y)$ is modeled as a multivariate Gaussian distribution.

Note: Since we're dealing with $p(\mathbf{x} | y)p(y)$, it is a generative model, despite its name!

- a. Consider two Gaussian distributions as formulated and visualized below:

$$\mathbf{x} | y = 0 \sim \mathcal{N}(\mu_{y=0}, \mathbf{I})$$

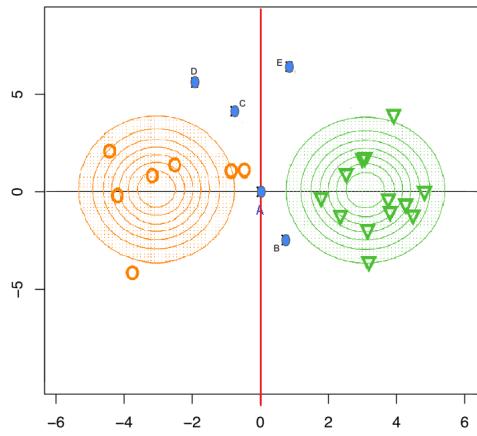
$$\mathbf{x} | y = 1 \sim \mathcal{N}(\mu_{y=1}, \mathbf{I})$$



- i. X are some observed data points. You are given that point A lies on the midpoint between the two distribution centers. Label each data point with its likely class. (hint: Both distributions have the same covariance!)

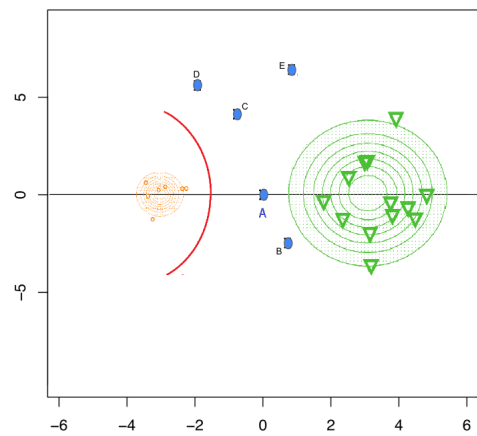
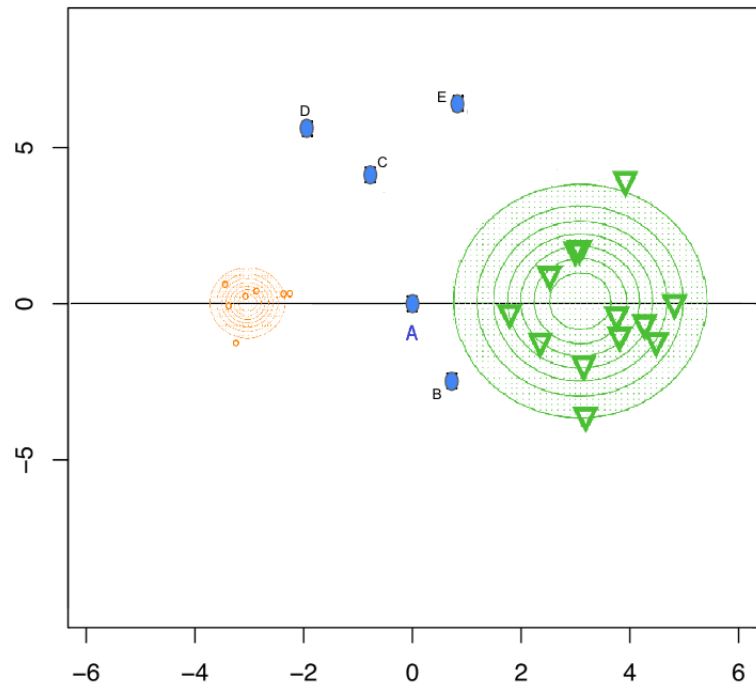
B - Green, C - Orange, D - Orange, E - Green

- ii. If we were to draw a boundary separating the two distributions, where would the boundary be and what would it look like?



The boundary will be linear.

- b. Now suppose a different distribution, whose covariance matrix is $\frac{1}{5}I$. Re-label the data points again. Point A lies on the midpoint between the two distribution centers. How does the boundary change?



The boundary will now be the zero-crossing of a quadratic function or an ellipse about the orange distribution (only a segment of it is drawn above).

7 Naive Bayes

7.1 Naive Bayes Recap

Naive Bayes is a probabilistic model that simplifies a generative model by using the naive Bayes assumption, which states that all input features of a data point are conditionally independent from each other given the output value. That is, for all $i \neq j$, X_i and X_j are conditionally independent given Y or:

$$P(X_1, X_2, \dots, X_j \mid Y) = \prod_{j=1}^M P(X_j \mid Y)$$

Naive Bayes Optimization:

1. Similar to MLE, estimate the parameters for the class prior distribution, $P(Y \mid \theta_{class})$.
2. **For each class label y** , use the data points with label y only to estimate the parameters (similar to MLE) for $P(X_j \mid Y = y, \theta_{\text{class conditional}, y, j})$ **for each feature X_j , independently**.

Naive Bayes Inference (Prediction): Given a new set of input features $x_1, \cdot, x_M$, class label with the highest posterior probability, $P(Y = y \mid x_1, \cdot, x_M)$ by taking advantage of Bays theorem:

$$\operatorname{argmax}_y P(Y = y) \prod_{j=1}^M P(X_j \mid Y = y)$$

7.2 Naive Bayes

Naive Bayes is a probabilistic algorithm that uses Bayes' theorem to classify data. Reminder that Bayes Theorem states

$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$

In addition to Bayes Theorem, Naive Bayes makes use of something called the Naive Bayes assumption, which states that features of a data point are conditionally independent given the class label. That is, all X_i are conditionally independent given Y or

$$P(X_1, X_2, \dots, X_i | Y) = \prod_{i=1}^I P(X_i | Y)$$

The **Naive Bayes Algorithm** works as follows:

1. Calculate the prior probability of each class label, $P(Y = y)$.
2. For each feature X_i of each class label, calculate the conditional probability $P(X_i | Y = y)$.
3. Given a new data point X , calculate the posterior probability of each class label using Bayes' theorem and the conditional probabilities from step (b). The class label with the highest posterior probability is the predicted class label.

In other words, assuming there are M features, the classifier can be written as

$$\operatorname{argmax}_y P(Y = y) \prod_{i=1}^M P(X_i | Y = y)$$

7.3 Laplace Smoothing

It's very convenient to use the Naive Bayes assumption that the features are independent given the class—it allows us to simplify the calculations a lot. But, whenever a word/token doesn't occur for a given class, the probability for that token given that class becomes 0, hence the probability of that class also becomes 0. This is a massive drawback. We cannot extrapolate our model to an example that contains a word that we have not seen before in our training data. If we tried, the probability for each class would just be 0.

Laplace smoothing is a technique used to address this issue by adding a small pseudo-count (denoted by α) to each observation. This ensures that no probability estimation becomes zero.

To be more precise,

$$P(x_i = k | Y = y) = \frac{(\text{Number of samples that belong to class } y \text{ and in which } x_i \text{ takes on class } k) + \alpha}{\text{Number of samples that belong to class } y + \alpha K}$$

K is the total number of classes that x_i can take on and α is a constant hyperparameter. Compare this to the formula from before:

$$P(x_i = k | Y = y) = \frac{\text{Number of samples that belong to class } y \text{ and in which } x_i \text{ takes on class } k}{\text{Number of samples that belong to class } y}$$

Note: Use $\alpha = 1$.

Note: It's important to clarify that "Laplace" smoothing is not related to the Laplace distribution. The term "Laplace" here simply refers to the technique of smoothing probabilities with pseudocounts.

Note: For the case of Bernoulli distributions, Laplace smoothing is equivalent to using a $\text{Beta}(\alpha + 1, \alpha + 1)$ prior. For categorical distributions, Laplace smoothing can be interpreted as using a Dirichlet prior.

Now, let's walk through an example where our goal is to classify whether an email is SPAM or not. Consider the following training samples:

SPAM	Email Body
1	Money is free now
0	Pat teach 380
0	Pat free to teach
1	Sir money to teach
1	Pat free money now
0	Teach 380 now
0	Pat to teach 380

The vocabulary consists of the following words: {380, free, is, money, now, Pat, Sir, teach, to, tomorrow}. Compute the following probabilities:

1. Fill in the tables below:

$P(Y = 1)$	$P(Y = 0)$

$$P(Y = 1) = 3/7 \quad P(Y = 0) = 4/7$$

j	$P(X_j = 1 \mid Y = 1)$	$P(X_j = 1 \mid Y = 0)$
380	1/5	4/6
free	3/5	2/6
is	2/5	1/6
money	4/5	1/6
now	3/5	2/6
Pat	2/5	4/6
Sir	2/5	1/6
teach	2/5	5/6
to	2/5	3/6
tomorrow	1/5	1/6

2. Consider the following email body: $X = \text{Pat teach now}$. Accounting for the laplace smoothing, fill in the table below:

j	x	$P(X_j = x \mid Y = 1)$	$P(X_j = x \mid Y = 0)$
380	0	$4/5$	$2/6$
free	0	$2/5$	$4/6$
is	0	$3/5$	$5/6$
money	0	$1/5$	$5/6$
now	1	$3/5$	$2/6$
Pat	1	$2/5$	$4/6$
Sir	0	$3/5$	$5/6$
teach	1	$2/5$	$5/6$
to	0	$3/5$	$3/6$
tomorrow	0	$4/5$	$5/6$

3. Reminder that with naive Bayes, $P(Y, X_1, \dots, X_M) = \prod P(X \mid Y)P(Y)$.

$P(Y = 1, X_1, \dots, X_M)$	$P(Y = 0, X_1, \dots, X_M)$

$$\begin{aligned}
 P(Y = 1, X_1, \dots, X_M) &= \prod P(X \mid Y = 1)P(Y = 1) \\
 &= \frac{4}{5} * \frac{2}{5} * \frac{3}{5} * \frac{1}{5} * \frac{3}{5} * \frac{2}{5} * \frac{3}{5} * \frac{2}{5} * \frac{3}{5} * \frac{4}{5} * \frac{3}{7} \\
 &= 0.00045500708
 \end{aligned}$$

$$\begin{aligned}
 P(Y = 0, X_1, \dots, X_M) &= \frac{2}{6} * \frac{4}{6} * \frac{5}{6} * \frac{5}{6} * \frac{2}{6} * \frac{4}{6} * \frac{5}{6} * \frac{5}{6} * \frac{3}{6} * \frac{3}{6} * \frac{4}{7} \\
 &= 0.00340213817
 \end{aligned}$$

	$P(Y = 1 \mid X_1, \dots, X_M)$	$P(Y = 0 \mid X_1, \dots, X_M)$
4.	$\frac{0.00045500708}{0.00340213817} = 0.118$	$\frac{0.00340213817}{0.00340213817} = 0.882$

We see that $P(Y = 1 \mid X_1, \dots, X_M)$ is no longer 0.

What's the effect of α ? Compare when $\alpha = 1$ with when $\alpha = \infty$

As α increases, the likelihood probability moves towards uniform distribution.