To cover:

- MLE vs. MAP
  - NB example
    - Priors: Beta \((B_H, B_L)\) vs. Uniform
  - Relation between prior and # of data points
- NB decision boundary
  - Linear vs. quadratic
  - Connection to Bayes optimal classifier
  - Multinomial & Dirichlet
  - Connection to Bernoulli and beta
- LR
  - Connection to naive bayes; be careful about conditions.
MLE vs. MAP

MLE and MAP are ways of estimating the parameters of a model.
MLR - maximum likelihood estimate (frequentist)
MAP - maximum a posteriori estimate (Bayesian)

Example: Flipping a coin:
\(X_n = \text{Bernoulli}(\theta)\), \(\theta\): probability coin comes up heads. Heads \(\theta = 1\), Tails \(\theta = 0\).

We are given a sample of \(n\) flips, \(X_1, \ldots, X^n\).

How do we guess the value of \(\theta\)?

MLE: choose \(\hat{\theta}_{\text{MLE}}\) to maximize the likelihood of \(X_1, \ldots, X^n\).

\[
\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \prod_{i=1}^{n} \theta^x_i (1-\theta)^{1-x_i}
\]

In class, we saw \(\hat{\theta}_{\text{MLE}} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\# \text{ of heads}}{\# \text{ of flips}}\).

A difference between MAP: Use Bayes' Rule to define a distribution over parameters \(\theta\) as a product of data likelihood and prior distribution.

Bayesians treat parameters (like \(\theta\)) as random variables.

Note that \(p(\theta | X_1, \ldots, X^n)\) is a distribution over possible parameters \(\theta\).

MAP estimate chooses the \(\theta\) which maximizes \(p(\theta | X_1, \ldots, X^n)\):

\[
\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} p(\theta | X_1, \ldots, X^n) = \arg \max_{\theta} \prod_{i=1}^{n} \theta^x_i (1-\theta)^{1-x_i} \cdot p(\theta)
\]
Issue: we must choose the prior $p(\theta)$.
Usually we choose conjugate priors which make MAP estimation easy. Conjugate of Bernoulli is Beta $\text{Beta}(B_H, B_T) = \frac{\theta^B_H (1 - \theta)^{B_T}}{B(\theta^B_H, \theta^{B_T})}$

So: $\hat{\theta}_{\text{MAP}} = \arg\max \theta \frac{\theta^{B_H-1} (1-\theta)^{B_T-1}}{B(\theta^B_H, \theta^{B_T})} B_H - 1 + \sum_i x_i (1-\theta)^{B_T - 1 + \sum_i x_i}$

And we see that $\hat{\theta}_{\text{MAP}} = \frac{\sum_i x_i}{n + B_H + B_T - 2}$; $B_H$ and $B_T$ act like "virtual" flips!

Bayesian Inference in Pictures:

In this picture, the dark line is the final posterior distribution. The light lines show what the posterior looks like as we observe more data points. Note that these distributions converge to the final posterior.

A common choice for $B_H, B_T$ is to set $B_H = B_T = \alpha$, in which case:

$\hat{\theta}_{\text{MAP}} = \frac{(\alpha - 1) + \sum_i x_i}{n + 2(\alpha - 1)}$; This looks like something on your problem set...
(Hint: set $\alpha = 2$)
**Multinomial & Dirichlet Distributions**

The multinomial distribution is like the Bernoulli distribution for more than 2 values.

- Bernoulli: flipping a coin (2 outcomes)
- Multinomial: rolling a die (6 outcomes)

Let's say we're rolling an unfair die. How many parameters do we need to describe the distribution? Answer: 5.

In general: a values means n-1 parameters.

θ is our parameter → θ_i is the probability of observing the i^th outcome. (e.g., θ_3 = probability of rolling 1 on the die)

Conjugate of multinomial distribution is the dirichlet distribution (like conjugate of Bernoulli is Beta)

**Note:**

- θ is a vector
- Dirichlet (α) = \( \frac{1}{B(\alpha)} \prod_{i=1}^{n} \alpha_i \)

As in Beta case, the θ_i values act like virtual observations!

\[ \hat{θ}_{PCE} = \frac{\text{Count}(X = i)}{\text{Total # of observations}} \]

\[ \hat{θ}_{MAP} = \frac{(d_i - 1) + \text{Count}(X = i)}{\text{Total # of observations}} + \sum_{i=1}^{n} (d_i - 1) \]
Naive Bayes Classifier

Let \( X = (X_1, \ldots, X_d)^T \) be a feature vector, \( Y \in \{0, 1\} \) be a label.

Recall: classification means predicting \( Y \) from \( X \), given training data
\[
D = \{(X_1', Y_1'), (X_2', Y_2'), \ldots, (X_n', Y_n')\}
\]

Naive Bayes Assumption: Probability of each feature \( X_i \)'s conditionally independent given \( Y \). Therefore, we can factor distribution \( P(Y, X) \)

\[
P(X, Y) = P(Y) \prod_{i=1}^{d} P(X_i | Y)
\]

class prior class conditional distributions

Class conditional distributions \( P(X_i | Y) \) may have many different forms. E.g.,
- \( X_i \) continuous - Gaussian
- \( X_i \) boolean - Bernoulli
- \( X_i \) categorical - multinomial

\[(MLE) \text{ Parameter Estimation: } \hat{\theta} = \frac{1}{n} \sum_{j=1}^{n} y_j\]

Let \( P(Y=1) = \pi \)

Estimate \( P(X_i | Y=y) \) based on the distribution.

Example: \( P(X_i | Y) \) is Gaussian with unit variance \( (\sigma^2 = 1) \)
\[
P(X_i | Y=y) = \frac{1}{\sqrt{2\pi}} \exp \left\{ - \frac{(x_i - \mu_i)^2}{2} \right\}
\]

\[
\hat{\mu}_i = \frac{1}{\sum_{j=1}^{n} (y_j = y)} \sum_{j=1}^{n} (y_j = y) X_i
\]

\[
\hat{\sigma}_i = \frac{1}{\sum_{j=1}^{n} (y_j = y)} \sum_{j=1}^{n} (y_j = y) (X_i - \hat{\mu}_i)^2
\]
Decision Boundaries

Consider the GNB classifier from the previous example. Assume we have two features \( X_1, X_2 \). The classifier will estimate one spherical Gaussian per class. Consider the level sets of these Gaussians:

\[ P(x_1, x_2 | y=1) \]

Shaded region is where the predicted class is 1 (i.e., any instance which falls in the shaded area will have predicted label 1).

Dashed line is the decision boundary, which divides the instance space based on the predicted class.

As another example, let's say we now estimate a standard deviation \( \sigma_{i,y} \) for each class conditional distribution:

\[ P(x_1, x_2 | y=0) \]

Notice that the level sets are now axis-aligned ellipses because the \( \sigma_{i,y} \)'s scale the dimensions differently.

In our first example, our decision boundary is linear, while in the second example it is a curve (actually a quadratic curve). The shape of a classifier's decision boundary is an important characteristic of the classifier.
Logistic Regression (LR):

LR model $P(Y|X)$ instead of $P(X,Y)$.

$$P(Y=1|X) = \frac{\exp \xi w_0 + \sum \xi w_i X_i}{1 + \exp \xi w_0 + \sum \xi w_i X_i}$$

N.B. $g(z) = \frac{\exp \xi z}{1 + \exp \xi z}$ is the logistic function.

Logistic function has a value between 0 and 1.
$g(z) \to 0$ as $z \to -\infty$
$g(z) \to 1$ as $z \to \infty$
$g(0) = 0.5$

LR's decision boundary is linear:

$$P(Y=1|X) = \frac{\exp \xi w_0 + \sum \xi w_i X_i}{1 + \exp \xi w_0 + \sum \xi w_i X_i} = \frac{\exp \xi w_0 + \sum \xi w_i X_i}{1 + \exp \xi w_0 + \sum \xi w_i X_i}$$

$$\Rightarrow 0 = w_0 + \sum \xi w_i X_i$$

linear equation for the decision boundary.

Where does the LR form come from?
Consider our Gaussian Naive Bayes classifier with unit variance for each class:

$$P(Y=0|X) = \frac{P(Y=0) P(X|Y=0)}{P(Y=0) P(X|Y=0) + P(Y=1) P(X|Y=1)}$$

$$= \frac{1}{1 + \frac{P(Y=1) P(X|Y=1)}{P(Y=0) P(X|Y=0)}}$$
Consider the second term in the denominators:

$$\frac{P(Y=1)P(X_1|Y=1)}{P(Y=0)P(X_1|Y=0)} = \exp \frac{\phi}{\epsilon} \ln \left( \frac{P(Y=1)}{P(Y=0)} \frac{P(X_1|Y=1)}{P(X_1|Y=0)} \right)$$

Our NB assumption says: $P(X_i|Y) = \prod_{j=1}^n P(X_j|Y)$

so:

$$\frac{P(Y=1)}{P(Y=0)} \frac{P(X_1|Y=1)}{P(X_1|Y=0)} = \exp \frac{\phi}{\epsilon} \ln \left( \frac{P(Y=1)}{P(Y=0)} \right) + \sum_{i=1}^n \ln \left( \frac{P(X_i|Y=1)}{P(X_i|Y=0)} \right)$$

Consider the terms in the sum:

$$\ln \left( \frac{P(X_i|Y=1)}{P(X_i|Y=0)} \right) = \ln \left( \frac{\exp \left( -\frac{\phi}{2} \left( X_i - \hat{\mu}_{i1} \right)^2 \right)}{\exp \left( -\frac{\phi}{2} \left( X_i - \hat{\mu}_{i0} \right)^2 \right)} \right)$$

$$= \ln \exp \left( -\frac{\phi}{2} \left( X_i - \hat{\mu}_{i0} \right)^2 \right) - \frac{1}{2} \left( X_i - \hat{\mu}_{i1} \right)^2$$

$$= \frac{1}{2} \left( X_i^2 - 2X_i \hat{\mu}_{i0} + \hat{\mu}_{i0}^2 \right) - \left( X_i^2 - 2X_i \hat{\mu}_{i1} + \hat{\mu}_{i1}^2 \right)$$

$$= (\hat{\mu}_{i1} - \hat{\mu}_{i0}) X_i + \frac{\hat{\mu}_{i0} - \hat{\mu}_{i1}^2}{2}$$

Note that the above equation is linear in $X_i$! So, if we set:

$$w_i = \hat{\mu}_{i1} - \hat{\mu}_{i0}$$

$$w_0 = \ln \left( \frac{P(Y=1)}{P(Y=0)} \right) + \sum_{i=1}^n \left( \frac{\hat{\mu}_{i0} - \hat{\mu}_{i1}}{2} \right)$$

Then:

$$P(Y=0|X) = \frac{1}{1 + \exp \phi w_0 + \sum_{i=1}^n w_i X_i}$$
WARNING: This derivation does not work for all types of NB classifiers. Example: our earlier GNB classifier with estimated per-class variances. (Recall the decision boundary was a quadratic curve, while L12's decision boundary is always linear!)

Multiclass Logistic Regression

What if we have $K$ labels, $Y_1, \ldots, Y_K$ instead of just 2 labels?

$$P(Y = y_j | X) = \frac{\exp \sum_{j} w_{0j} + \sum_{i=1}^{d} w_{ij} X_i}{1 + \sum_{k=1}^{K} \exp \sum_{j} w_{0k} + \sum_{i=1}^{d} w_{ik} X_i}$$

$$P(Y = y_k | X) = \frac{1}{1 + \sum_{j=1}^{K} \exp \sum_{j} w_{0k} + \sum_{i=1}^{d} w_{ik} X_i}$$

Decision Rule: choose $Y_k$ with maximum probability.

$$\hat{Y} = \text{arg max}_Y P(Y = y | X)$$

$$= \text{arg max}_j \left( 0, w_{0j} + \sum_{i=1}^{d} w_{ij} X_i \right)$$

Ex. 2 classes decision boundary

Dashed line $\hat{y}$:

$$w_{01} - w_{02} + \sum_{i=1}^{d} (w_{1i} - w_{2i}) X_i = 0$$

(can predict 2)

$$w_{01} + \sum_{i=1}^{d} w_{1i} X_i > 0$$

(can predict 1)